IDENTIFYING INTERACTIONS CONTROLLING PROTEIN STRUCTURE AND TRANSITION KINETICS VIA EFFICIENT SIMULATION METHODS

by Camilo Andres Velez Vega

IDENTIFYING INTERACTIONS CONTROLLING PROTEIN STRUCTURE AND TRANSITION KINETICS VIA EFFICIENT SIMULATION METHODS

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

In Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Camilo Andres Velez Vega

August 2010

IDENTIFYING INTERACTIONS CONTROLLING PROTEIN STRUCTURE AND TRANSITION KINETICS VIA EFFICIENT SIMULATION METHODS

Camilo Andres Velez Vega, Ph. D.

Cornell University 2010

Advances in hardware and molecular force fields have given a boost to computational studies of the thermodynamics and dynamics of transitions involving small protein systems. Because these studies are still hampered by the vast amount of CPU time required, new approaches and sampling optimization strategies are still needed.

In this work, several schemes were developed and used to increase the simulation efficiency of various proteins and give insights on their structure, kinetics and mechanism. Our studies focus on the recognition of particular markers that assist in antibody design or lead to protein misfolding and aggregation. Concerning structural identification, the application of novel techniques based on the Replica Exchange Method is illustrated by a mutagenesis analysis seeking to “humanize” the hypervariable regions of a llama heavy-chain antibody. Regarding kinetic and mechanistic characterization, the application of optimization schemes of the Forward Flux Sampling Method is demonstrated by the study of structural transitions for the Alanine Dipeptide and the Tryptophan Cage synthetic protein. Our results are in good agreement with previous experimental studies performed on these systems and further characterize the pathways and “transition states” traversed in the studied transitions.

BIOGRAPHICAL SKETCH

Camilo Velez Vega was born Bogotá, Colombia. In 1995 he started his studies as a Chemical Engineer in Universidad Nacional de Colombia, obtaining his Bachelor's Degree in 2000. Shortly after, he pursued a Master of Science in Chemical Engineering at Texas A&M University, working for Dr. Nikolaos Kazantzis and Dr. Theresa Good, and obtained his Degree in 2002. He was then hired as a Process Engineer in Refineria de Nare S.A., located in Puerto Perales, Colombia, job that he held for one year (July 2002-June 2003). Subsequently, he was promoted to Technical Director in the same company, and worked in this position for almost two years. In August 2005 he joined the Ph.D. program of the Chemical and Biomolecular Engineering Department at Cornell University, mentored by Dr. Fernando Escobedo.

Dedicado a mi familia, a mi mentor y a mis amigos cercanos, quienes me han apoyado incondicionalmente en todas las decisiones tanto personales como profesionales que he tomado durante mi Doctorado.

ACKNOWLEDGMENTS

I gratefully acknowledge the valuable and unconditional guidance that my advisor Dr. Fernando Escobedo has given me throughout my Ph.D. studies. I have been privileged to have him as my mentor, and his lessons have greatly promoted my growth as a professional.

TABLE OF CONTENTS

BIOGRAPHICAL SKETCH.....iii

DEDICATION.....iv

ACKNOWLEDGMENTS.....v

TABLE OF CONTENTS.....vi

LIST OF FIGURES.....vii

LIST OF TABLES.....ix

CHAPTER 1. Simulated Mutagenesis of the Hypervariable Loops of a Llama VHH Domain for the Recovery of Canonical Conformations.....1

CHAPTER 2. Kinetics and Reaction Coordinate for the Isomerization of Alanine Dipeptide by a Forward Flux Sampling Protocol.....39

CHAPTER 3. Kinetics and Mechanism of the Unfolding N-L Transition of Trp-cage in Explicit Solvent via Optimized Forward Flux Sampling Simulations.....79

CHAPTER 4. Correlating Structural and Solubility Behavior of Selected Aβ-42 Polypeptide Mutants……...115

APPENDIX A.....146

APPENDIX B.....150

APPENDIX C. Comparative Analysis of BG Protocols.....154

APPENDIX D. Estimation of the Rate Constant for Initial Set of FFS Runs.....160

LIST OF FIGURES

Figure 1.1. Probabilities of occurrence for the RMSD values of the loops of the crystal (*1HCV*) and twenty NMR (*VHH-H14*) structures relative to reference structures 1DFB and 1FVC.....6
Figure 1.2. Probabilities of occurrence for the RMSD values of the simulated wildtype llama VHH H1 (*A*) and H2 (*B*), and H3 (*C*) loops at 300°K relative to reference structure 1DFB and 1HCV.....12
Figure 1.3. RMSD values between the conformations of the simulated H1 loop and 1DFB, 1HCV (*A*), and the simulated H2 loop and 1DFB, 1FVC (*B*), of a 10-ns run for the 3-FFSa case.....19
Figure 1.4. Simulated H1 (*residues 26-32*) and H2 (*residues 52,52a-56*) loops from the low RMSD structures of mutants 3-FFSa, 2-FL, and 1-Fa, showing the H-bonding pattern for each case.....20
Figure 1.5. Simulated loops of a representative 3-FFSa structure, as compared to the initial 3-FFSa conformation used for the simulation and the reference structures 1DFB (*H1 loop*), 1FVC (*H2 loop*), and 1HCV (*H3 loop*).....21
Figure 1.6. Representative structure of the simulated H1 loop of mutant 2-FS as compared to 1DFB H1 loop (*A*) and a low RMSD H1 loop of mutant 3-FFSb (*B*).....22
Figure 1.7. Walk over temperature space for the data evaluation period of a 6-mutant MMREM run. The two replicas for mutants 3-FFSa (*A*) and 3-FFSb (*B*) are displayed in each plot.....27
Figure 1.8. RMSD values between the conformations of the simulated H1 loop and 1DFB, 1HCV (*A*), and the simulated H2 loop and 1DFB, 1FVC (*B*), of a 10-ns run for the 3-FFD case.....28
Figure 1.9. Simulated H1 (*residues 26-32*) and H2 (*residues 52,52a-56*) loops of a representative 3-FFD structure, showing its H-bonding pattern.....29
Figure 2.1. A model for alanine dipeptide.....41
Figure 2.2. A schematic view of the generation of branched paths (thick lines) using the branched growth (BG) sampling method.....44
Figure 2.3. (*A*) Distribution for the center of mass velocity of water molecules for MD simulations using: (–) thermostat A and (•) Nosé-Hoover thermostat. (*B*) Time progression of the velocity autocorrelation functions (VACF) for thermostat A (–) and (•) Nosé-Hoover thermostat.....51
Figure 2.4. Free energy landscape for blocked alanine dipeptide in vacuum at 300 K 54
Figure 2.5. Free energy landscape for alanine dipeptide in explicit solvent at 300 K 58
Figure 2.6. Results for the FFS-MC simulations in vacuum at 300 K.....65
Figure 2.7. Results for the FFS-MD simulations in vacuum at 300 K.....66
Figure 2.8. Density map (P_{TSE}) obtained from the TPE for several FFS-MD runs for the $C5/C7_{eq} \Rightarrow \beta_2/\alpha_R$ reaction at 300K in explicit solvent.....71
Figure 2.9. Results for the optimization process of the λ_0 positioning in the FFS-MD simulation for the $C5/C7_{eq} \Rightarrow \beta_2/\alpha_R$ reaction at 300K in explicit solvent.....72

Figure 2.10. Isocommittor surfaces obtained during the FFS-MD simulations for the $\beta_2/\alpha_R \Rightarrow C5/C7_{eq}$ reaction at 300K in explicit solvent.....74
Figure 3.1. ACF(lag) for states at λ_0 =0.06 nm as a function of the separation between stored states (lag), for y = $RMSD_{ca}$ and y = n_{wat}.....96
Figure 3.2. Optimization of the λ_0^{opt1} positioning.....97
Figure 3.3. Distribution of conformations at $RMSD_{hx} = \lambda_0$ (red) used as starting states for the FFS runs, juxtaposed over the phase space sampling of basin A.....98
Figure 3.4. Isocommittor surfaces of the reaction coordinate model for the N-L transition.....102
Figure 3.5. Plots of $RMSD_{hx}$ vs. λ^{opt1} (*A*) and λ^{opt2} (*B*) in the region of attraction of basin A.....103
Figure 3.6. Representative structures of conformations at λ_A^{opt2} (left), TSE (middle) and λ_B^{opt2} (right).....104
Figure 3.7. Optimized reaction coordinate model vs. various order parameters along the TPE.....108
Figure 4.1. Average Backbone RMSF for the monomers studied…….....123
Figure 4.2. Hydrophobic Solvent Accessible Surface Area of the monomers studied, normalized by the number of residues in each segment……..124
Figure 4.3. Relative population of the clusters identified for the Dutch (*A*), WT (*B*) and GM6 (*C*) ensembles at 296 K, for the 15-25 ns period……...127
Figure 4.4. Contact maps for the representative structures of the three major clusters of the Dutch (*A*) and WT (*B*) monomers, and the two major clusters of GM6 (*C*)…….129
Figure A.1. Probabilities of occurrence for the RMSD values of the simulated mutants 2-FF and 3-FFSa for H1 (*A*) and H2 (*B*) loops at 300°K relative to reference structures 1DFB and 1FVC.....149
Figure C.1. Contour graph of the free energy surface for the two-dimensional potential.....158
Figure C.2. Ratios between the rate constant found for different BG schemes and the one obtained from Brute Force simulations for CBG (black bars), original BG (grey bars) and RBG (white bars) schemes.....159
Figure D.1. Sampling of the TPE obtained from our initial set of FFS runs.....163

LIST OF TABLES

Table 1.1. Mutations for the different cases simulated using conventional REM.....14
Table 1.2. Percentage of loop conformations with RMSD values lower than 1.2 Å with respect to reference structures, for the data evaluation period of conventional REM simulations.....14
Table 1.3. Average of the RMSD values with respect to the typical structure, for the data evaluation period of conventional REM simulations.....14
Table 1.4. Initial temperature distribution and average between the most visited temperatures for the two replicas of each mutant, for the data evaluation period of a 10-ns 6-mutant MMREM run.....26
Table 1.5. Initial temperature distribution and most visited temperature for each mutant, for the data evaluation period of a 10-ns 12-mutant MMREM run.....26
Table 2.1. Optimized move set for MC simulation in vacuum.....57
Table 2.2. Optimized {λ} sets for vacuum and explicit solvent FFS-MC and FFS-MD simulations.....57
Table 2.3. LSE parameters and analysis of variance for the reaction coordinate model of the FFS-MC simulation in vacuum.....63
Table 2.4. LSE parameters and analysis of variance for the reaction coordinate model of the FFS-MD simulation in vacuum.....63
Table 2.5. LSE parameters and analysis of variance for the reaction coordinate model of the slower $C5/C7_{eq} \Rightarrow \beta_2/\alpha_R$ reaction of an FFS-MD simulation in explicit solvent 69
Table 2.6. LSE parameters and analysis of variance for the reaction coordinate model of the faster $\beta_2/\alpha_R \Rightarrow C5/C7_{eq}$ reaction from a FFS-MD simulation in explicit solvent 73
Table 3.1. Initial, optimized and reference $\{\lambda\}$ sets for order parameter $\lambda = RMSD_{hx}$..94
Table 3.2. Initial and optimized $\{\lambda'\}$ sets for the reaction coordinate model λ^{opt1}.....95

CHAPTER 1

SIMULATED MUTAGENESIS OF THE HYPERVARIABLE LOOPS OF A LLAMA VHH DOMAIN FOR THE RECOVERY OF CANONICAL CONFORMATIONS*

I. INTRODUCTION

The chemical diversity of the antigen binding sites of antibodies plays a critical role in the immune system by facilitating the identification of countless foreign antigens. This diversity is due largely to the amino acid variability of the complementarity-determining regions (CDRs). Structurally, the binding sites of conventional antibodies are composed of three light chain and three heavy chain hypervariable regions (denoted L1, L2, L3, H1, H2, and H3, respectively), which represent six loops that pack together to form a versatile surface for molecular recognition. Accordingly, the problem of antigen binding site engineering is often simplified to the modeling and structure prediction of a set of six loops on a relatively conserved protein scaffold.

A variety of knowledge-based (KB),[1-19] ab-initio,[20-33] and combined methods[34-36] have been used for predicting the structures of antibody hypervariable loops (for a comprehesive review of solutions for the more general problem of loop structure prediction for proteins see, e.g., Fiser et al.[37]). In particular, the use of KB methods has shown promising results[2,11] due to the limited number of conformations that have been observed for most antibody loop backbones. Al-Lazikani et al.[1] and Martin et al.[11] have categorized these loops into various canonical conformations based on available crystal structures. Shirai et al.[18] proposed a more rigorous classification for H3 loops.

Despite the successful effort of grouping many of the experimentally resolved hypervariable structures, several loops do not fit within these classifications. Such is the case for some of the loops of the variable domains of camelid heavy chain antibodies (VHHs), which lack light chains.[5,38] These antibodies utilize a reduced biomolecular surface for antigen binding which appears to have evolved increased hypervariable loop structural variability to compensate for the absence of the light chain.

Camelid VHHs have increasingly been used for engineered antigen binding given their various desired characteristics, which include small size, high expression level,[39] reversible folding after exposure to high temperatures,[40] and functionality as enzyme inhibitors.[41] Nevertheless, their general application still faces some limitations, such as changes in structure upon loop grafting or exposure to harsh conditions like increased temperature and low pH.[42] A computational method that can assist in the identification of key mutations that lead to affine structures with improved stability under different conditions is highly desirable.

A particular VHH amenable for systematic study of its key loop residues is the llama VHH raised against human chorionic gonadotropin (hCG). Its antigen-free structure has been resolved by Renisio et. al.[43] via NMR spectroscopy (PDB code 1G9E and referred to as VHH-H14) and by Spinelli et. al.[44] via X-ray crystallography (PDB code 1HCV). Whereas the crystal structure shows well-defined canonical class-2A H2 and non-canonical H1 conformations, the NMR structure ensemble of these loops is more indistinct and suggests significant flexibility in solution. Notably, KB methods provide either ambiguous or incorrect predictions for each of the three loops of the crystal structure, motivating the use of more sophisticated methods for structure prediction and computer analysis. For example, using the hybrid Monte Carlo replica exchange (HYMREX)[22] method, the crystal structure loop conformations were

simulated and examples of loop flexibility were provided that are reflected in the dynamic equilibrium observed using NMR. Although prior studies of this system have suggested several residues to be key determinants of the loop conformations,[5,22,44] additional insights may be gained from site-directed mutagenesis.

In the present study, replica exchange molecular dynamics (REM) simulations of wildtype and mutant llama VHH hypervariable loops were performed to quantify loop flexibility and identify residues that play a critical role in shaping and stabilizing these conformations. Much of the analysis is focused on the central H1 loop, which contains predominantly hydrophilic residues and adopts a backbone structure that is unique among the known H1 conformations in antibodies.[5] In contrast, the vast majority of H1 loops found in conventional antibodies adopt a well-defined conformation that is kinked in the center and presumably stabilized by key hydrophobic residues. One of our principal aims was to identify an effective strategy and a solvent force field for simulating the conformational equilibrium of such canonical and non-canonical antibody loops, and for identifying mutations that can account for their differing backbone structures. An All Pairs Exchange[45] adaptation of REM[46] was implemented in parallel using CHARMM[47] version 32 via the Multiscale Modeling Tools for Structural Biology (MMTSB) toolset.[48] Furthermore, a novel application of REM for rapid mutant screening was implemented, in which various plausible mutations were evaluated in a single REM run.

This work is organized as follows: The crystal and NMR structures of the wildtype llama VHH hypervariable loops are reviewed first, followed by a brief analysis of the mutations selected and a description of the simulation details. The simulated conformational changes of the loops are then compared with those observed in experimental structures. Finally, the simulation results on the wildtype structure and the various mutants considered are presented and discussed in the context of

stabilization of canonical structures. The paper ends with some concluding remarks about the structural determinants and the simulation method.

II. LLAMA VHH LOOP CONFORMATIONS

The available NMR[43] and crystal[44] structures of the anti-hCG llama VHH domain allow the identification of some key structural features of the hypervariable loops. For this study, H1, H2, and H3 comprise 1HCV residues 26-32 (GRTGSTY), 52,52a-56 (NWDSAR), and 95-102 (GEGGTWDS), respectively. Fig. 1.1 shows the relative probabilities of occurrence for the RMSD values between the loops of the crystal structure and those of the twenty NMR structures reported (unless otherwise indicated, RMSD refers to backbone atom RMSD). For this calculation, each loop from the NMR structure ensemble was aligned to the corresponding loop of the crystal conformation. Noting that the crystal structure was obtained at lower temperatures with crystal packing whereas the NMR spectra were recorded at 300°K in solution, some discrepancy may be expected.

For the H1 loop, Fig. 1.1A illustrates NMR configurations within a range of 1.4-2.2 angstroms (Å) from the crystal structure, the representative NMR structure deviating by 1.4 Å. In addition, Fig. 1.1B shows the RMSD distribution for the NMR H1 loops relative to those of the crystal structure of a monoclonal antibody Fab fragment (PDB code 1DFB).[49] 1DFB adopts a type 1 H1 loop conformation and is therefore used as reference. Although the KB methods of Al-Lazikani et al.[1] and Martin et al.[11] predict a type 1 conformation for the H1 loop of the llama VHH, it is evident that no NMR conformation adopts a type 1 structure. This is also true for the crystal H1 loop, which has a RMSD value of 1.4 Å with respect to the corresponding 1DFB loop.

For the H2 loop, it can be seen in Fig. 1.1A that the NMR structures are within a range of 0.8-1.8 Å from the crystal structure and the representative NMR structure differs by 1.6 Å. Comparisons are made with the H2 loop structures of two reference antibodies (Fig. 1.1C), namely, 1DFB and the FV fragment of the humanized antibody 4D5[50] (PDB code 1FVC), which adopt a type 3 and a type 2A conformation, respectively. KB predictions for the llama VHH H2 loop indicate either a type 3[1] or a type 2A[11] conformation. Fig. 1.1C shows that 45% and 25% of the conformations have RMSD values below 1.5 Å with respect to a type 3 and a type 2A structure, respectively. Conversely, the RMSD values of the 1HCV H2 loop with respect to the type 3 and type 2A reference structures are 1.8 and 0.5 Å, respectively. It can then be inferred that, even though H2 is markedly type 2A in the crystal structure, it may be flexible in solution and temporarily adopt a type 3 structure. However, it is noted that residues D53 and S54 were not assigned in the NMR spectra.

Finally, for the H3 loop, Fig. 1.1A shows consistency between the crystal and NMR structures. This is indicative of an H3 conformation with reduced flexibility. All of the H3 structures adopt a kinked conformation, in contrast to the extended structure predicted using the H3 KB rules.[22]

III. SELECTION OF CANDIDATE MUTANTS

Simulated site-directed mutagenesis experiments were performed in an attempt to return the llama VHH H1 loop structure to the type 1 conformation predicted by KB methods. Only sites within the H1 and H2 regions that are thought to directly shape the standard H1 conformation were mutated. As first noted by Spinelli et al.,[44] the 1HCV H1 amino acid sequence coincides with a type 1 loop in four of the seven residues. The three residues considered to be infrequent were R27, G29 and T31. T31

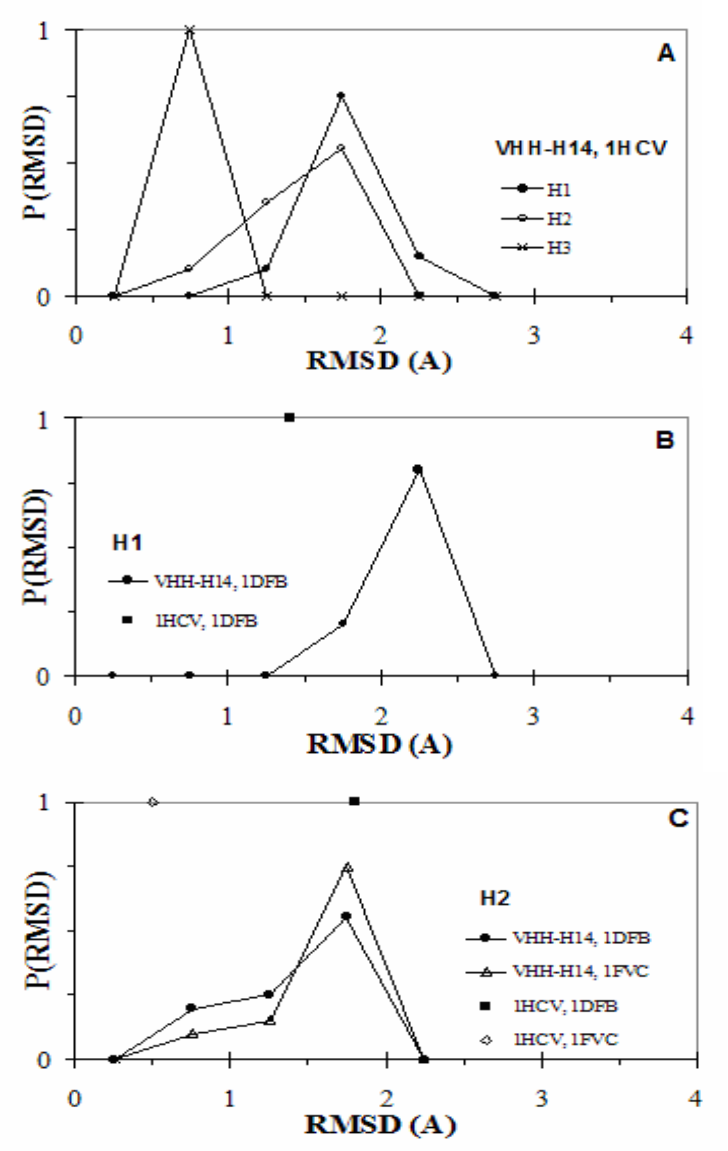


Figure 1.1. Probabilities of occurrence for the RMSD values of the loops of the crystal (*1HCV*) and twenty NMR (*VHH-H14*) structures relative to reference structures 1DFB and 1FVC. The plots show values for: the VHH-H14 structures with respect to 1HCV for H1, H2, and H3 loops (*A*); the VHH-H14 and crystal structures with respect to 1DFB for the H1 loop (*B*); and the VHH-H14 and crystal structures with respect to 1DFB and 1FVC for the H2 loop (*C*). For every case, zero probability indicates that the loops did not sample conformations having such RMSD.

is observed in 10.6% (33/310) of the H1 loops of antibody structures analyzed, whereas R27 and G29 are found in less than 1% of the cases. Given the moderate likelihood of finding T31 in H1 type 1 structures, our study focused on positions 27 and 29 as key determinants of the structural features. Moreover, residue W52a in H2 was identified as being important for preventing the type 1 H1 structure formation, due to possible clashing with residues in this canonical form ("a" in W52a indicates that this residue is between landmark alignment positions N52 and D53). Residues R27, G29, and W52a were consequently selected as targets for mutation; residue N52 was also selected as a possible mutation site for reasons that are clarified below. Single point mutations at position 27 or 29 as well as double point mutations at these two locations were then assumed to be the ones likely to achieve the desired conformation when considering changes in the H1 loop alone. Three and four point mutations included changes in sites from the H2 loop.

The substitute residues for sites 27 and 29 were chosen on the basis of frequent occurrences (Y/F/G at position 27 and F/I/L at position 29) in canonical type 1 H1 loops.[11] Incidentally, mutations in these sites have been proposed as an alternative way for achieving an augmented affinity in VHHs.[51] For most of the simulations involving mutations at either or both of these positions, we selected F because (with human therapeutic applications in mind) it occurs at these positions in most of the approximately 51 human germline VH gene segments (26 and 33 of 51 for positions 27 and 29, respectively).[52] Furthermore, we observed that its side chain was easier to bury than that of Y at position 27 (perhaps due to its higher hydropathy index). As for mutations in H2, a W52aS replacement was simulated to examine the steric hindrance effect of this bulky hydrophobic residue on the attainability of a type 1 H1 conformation. The serine residue at this position was selected for its relatively small size and to avoid an evident bias toward an H2 type 3 (common residues D/P) or H2

type 2A (common residues P/T/A) conformation. Finally, a N52S mutation was simulated to evaluate the possible cooperative adoption of type 1 H1 and type 3 H2 loops, as occurs in antibody 1DFB, which also has a tryptophan at position 52a.

IV. SIMULATION METHOD

A reduced model of the crystal structure (1HCV) previously defined[22] was used as the basis for all simulations. Briefly, the hypervariable regions and a subset of residues in the proximal framework regions of 1HCV were chosen for our simulations. To reduce computation time, distal framework residues that are unlikely to have a major influence on the structure of the hypervariable loops were excluded. Altogether, 57 of the 117 amino acids in 1HCV were included in our simulations. C_{α}, N, and C backbone atoms that do not belong to hypervariable loops were restrained with a harmonic constant K_{HARM} of 0.5×(atomic mass), while backbone atoms from hypervariable regions and side chains from all residues included in the simulations were free to move. These restraints were imposed to simulate the limited motion displayed by backbone atoms of framework residues that are part of secondary (β-sheet) structures. Acetyl and N-methyl groups were used to cap the N- and C- terminal ends of each of the simulated fragments. The starting structure was slowly heated to 300°K via MD. Point mutations were introduced into the heated structure to realize the starting point for each of the mutant cases indicated in Table 1.5; this was followed by 1000 steps of steepest descent energy minimization and a short equilibration. The resulting conformation was the initial structure for each mutant. The velocity Verlet algorithm was chosen as the integrator with a time step of 2 fs. SHAKE[53] was used to constrain the lengths of bonds involving hydrogen atoms. Nonbonded interactions were calculated as described by Brooks et al.,[47] with a cutoff for the non-bonded list generation of 20 Å**,** a cutoff for non-bonded interactions of 18 Å, and an onset of the

switching function for non-bonded interactions of 16 Å**.** Other parameters used for REM are the default ones included in the MMTSB toolset.[48]

As mentioned in the introduction section, REM is a useful technique that has been applied to a wide variety of systems.[54] In general, M replicas of the system are simulated at M temperatures, and configurations are periodically exchanged in accordance with the Metropolis criterion. In this way, it is possible to sample large portions of phase space at high temperatures, producing structural changes than can favor visiting more constrained regions of phase space at low temperatures in an efficient manner. The configurations obtained at the lower temperatures of interest therefore reflect an enhanced sampling that would be very difficult to achieve using standard MD, even with a simulation time orders of magnitude longer. The All Pairs Exchange[45] variation enhances efficiency by redefining the generation probability in such a way that all possible replica pairs become candidates for exchange.

Various preliminary validation runs performed using CHARMM22-CMAP[55] with GBSW implicit solvent,[56] and CHARMM19[57] with either GBMVA[58] or EEF[59] implicit solvent models, are reviewed in Appendix A. It is noted that a solvent accessible surface area (SASA) calculation is included in the CHARMM GBMVA module. The results of these simulations suggested the implementation of REM via the CHARMM19 force field with GBMVA. Interestingly, Olson et al.[60] have recently observed that CHARMM19 may in certain cases be more appropriate for loop modeling with implicit solvent than newer force fields such as CHARMM22. All of the REM simulations discussed in this work consisted of 12 replicas spanning a temperature range of 300-900°K. Swaps between temperatures were attempted every 500 MD steps, and configurations were stored with the same frequency. Individual temperatures were maintained using the Nose-Hoover thermostat. Two approaches were tested for obtaining REM temperatures that enhance equilibration of the entire

system. In the first method, temperatures were chosen such that an acceptance ratio of approximately 30% was achieved for swap moves.[48,61,62] The second approach was that proposed by Trebst et al.,[63] which aims at the maximization of the number of round trips of replicas moving between the lowest and highest temperatures. Since no substantial improvement was observed when using the latter approach, the simulations reported below correspond to those obtained via the first technique.

V. RESULTS AND DISCUSSION

Wildtype VHH hypervariable regions

The structural variability of the loops for the wildtype system was assessed via a 10-ns REM simulation. The energy minimized model of 1HCV was used as the initial conformation as well as the reference for the calculation of loop RMSD. Its RMSD values from the crystal H1, H2 and H3 loops are 0.91 Å, 0.29 Å and 1.12 Å, respectively. The first 5 ns was considered to be an equilibration period; the results reported below correspond to the last 5 ns of the run, referred to as the data evaluation period.

Fig. 1.2A shows the RMSD distribution for the simulated H1 loop at 300°K relative to 1DFB (H1 type 1) and 1HCV. Sampling of the crystal structure is evident, with 35% of the configurations falling below 1.2 Å from the 1HCV H1 loop. Notably, the Y32 side chain of the low RMSD conformers lies flat under the loop, positioned analogously to that observed in the crystal structure. In contrast, the sub-ensemble of structures corresponding to RMSD values of around 1.8 Å have their Y32 side chain exposed to the solvent, oriented towards the H3 loop. These higher RMSD values may further be caused by a kink observed in the backbone of residue 28, which is stabilized by an H-bond (absent in the crystal structure) between the side chains of T28 and Q3. Overall, the structures obtained lie within a range of 0.5-2.5 Å from 1HCV, supporting

the variability of this loop observed in the NMR study. None of the conformations achieve H1 RMSD values below 1.2 Å from a type 1 loop, supporting the non-canonical nature of this loop.

Fig. 1.2B shows the RMSD probabilities for the H2 loop. The simulated conformations adopted by this loop are compared to the canonical type 3 (1DFB) and to the 1HCV loop (type 2A). None have RMSD values that fall below 1.2 Å from a type 3 loop, while 99% have RMSD values that are under this threshold when compared to the 1HCV loop. All of the H2 conformations display the N52-R56 H-bond observed in type 2A loops.

As evidenced from Fig. 1.2C, the H3 loop structures obtained by simulation display moderate structural variation, with structures between 0.7-2 Å with respect to the 1HCV structure. These conformations resemble the kinked shaped NMR and crystal H3 loops and have two corresponding E96/T99 intraloop backbone-backbone H-bonds (not shown). However, an increased flexibility is observed for these conformers when compared to the NMR structures, perhaps due to model deficiencies in solvent mediated stabilizing interactions within the H3 loop.

Mutational Analysis

In an effort to identify key mutations of the llama VHH that lead to an H1 loop with a type 1 conformation, different mutants containing anywhere from 1 to 3 point mutations were simulated for 10 ns using conventional REM (see Table 1.1). The initial structures from which the mutants were simulated have RMSD values ranging from 0.93-1.05 Å, 0.41-0.47 Å, and 0.87-0.99 Å from the H1, H2, and H3 loops, respectively, of 1HCV. With respect to their corresponding canonicals, these initial structures have RMSD values of 1.63-1.78 Å from 1DFB for the H1 loop, and 0.39-0.49 Å from 1FVC for the H2 loop. Each mutant was simulated for 5 ns (equilibration

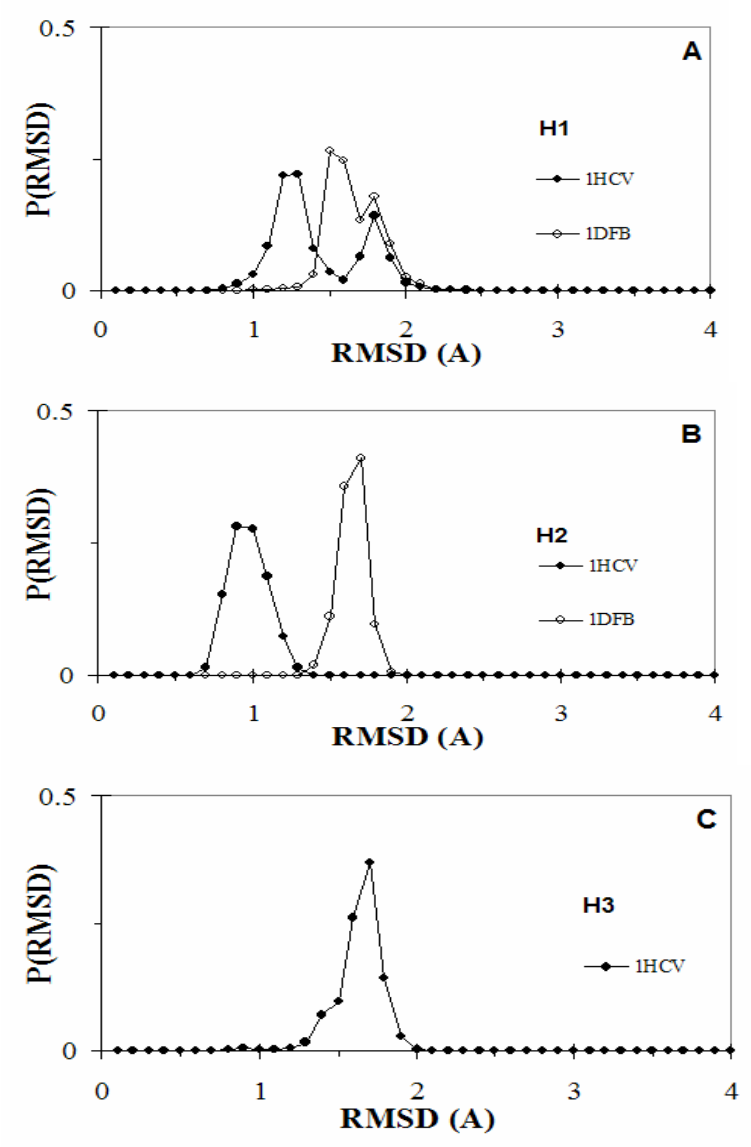


Figure 1.2. Probabilities of occurrence for the RMSD values of the simulated wildtype llama VHH H1 (*A*) and H2 (*B*), and H3 (*C*) loops at 300°K relative to reference structure 1DFB and 1HCV. The distributions account for the data evaluation period of a 10-ns run.

period), and then assayed at 300°K during the following 5 ns (data evaluation period) to detect signs of a stable type 1 H1 structure. Two markers used for this purpose were (i) the percentage of configurations with H1 RMSD values lower than 1.2 Å with respect to 1DFB (Table 1.2), and (ii) the average RMSD of each of the three loops with respect to the "typical" simulated structure, which corresponds to the structure that is closest in RMSD to the mean RMSD of the simulated H1 conformations from the canonical 1DFB loop (Table 1.3). It was verified that each such typical structure was consistent with the structures from the most frequent ensemble observed for each mutant. It is noted that even though the experimentally observed canonical loops are expected to be highly stable, non-canonical stable structures may also be obtained through mutagenesis. Thus, the second marker is used here as an appropriate indicator of the structural stability of each mutant.

Tables 1.2 and 1.3 show that the 3-FFSa conformers adopt a stable H1 type 1 structure. 99% of the H1 configurations have a RMSD value lower than 1.2 Å from 1DFB (Fig. 1.3A), and a 0.55 Å average RMSD from the typical structure is obtained. Conversely, for the 2-FF case, despite having two key mutations at positions 27 and 29, thought to be sufficient to reshape the non-canonical H1 loop back to a canonical type 1 structure,[1,11] only 70% of the configurations are below the selected threshold. The results for these two cases highlight the importance of interloop interactions on the stabilization of the simulated H1 type 1 conformation (plots of the RMSD distributions at 300°K for 2-FF and 3-FFSa are given in Fig. A.1, Appendix A).

Likewise, the 2-FL and 1-Fa mutants achieve a modest number of conformations with low values of RMSD from 1DFB. Visual inspection of the low RMSD configurations for the 3-FFSa, 2-FL and 1-Fa mutants confirms that type 1 H1 structures similar to the ones observed experimentally are obtained (Fig. 1.4). For 3-

Table 1.1. Mutations for the different cases simulated using conventional REM.

Case	Mutations
Wildtype	-
1-Fa	R27F
1-Fb	G29F
2-FF	R27F , G29F
2-FS	G29F , N52S
2-FL	R27F , G29L
3-FFSa	R27F , G29F , W52aS
3-FFSb	R27F , G29F , N52S

Table 1.2. Percentage of loop conformations with RMSD values lower than 1.2 Å with respect to reference structures, for the data evaluation period of conventional REM simulations.

Case	H1 (*ref. 1DFB*)	H2 (*ref. 1FVC*)
Wildtype	0%	85%
1-Fa	10%	65%
1-Fb	11%	95%
2-FF	70%	96%
2-FS	6%	99%
2-FL	15%	89%
3-FFSa	99%	97%
3-FFSb	29%	86%

Table 1.3. Average of the RMSD values with respect to the typical structure, for the data evaluation period of conventional REM simulations.

Case	H1 (Å)	H2 (Å)	H3(Å)	Average for the three loops (Å)
Wildtype	1.12	0.45	0.59	0.72
1-Fa	1.21	0.51	0.49	0.74
1-Fb	0.67	0.57	0.47	0.57
2-FF	0.51	0.58	0.49	0.53
2-FS	0.56	0.47	0.50	0.51
2-FL	1.05	0.49	0.53	0.69
3-FFSa	0.55	0.47	0.52	0.51
3-FFSb	1.29	0.58	0.73	0.87

FFSa**,** three H1 intraloop H-bonds are detected at the lowest temperature and coincide with those present in the H1 loop of the 1DFB crystal structure. However, this is in clear contrast with the H1 wildtype crystal structure, for which a single interloop H-bond is observed between G29 and Y32. The flexible nature of the wildtype conformer as compared to 3-FFSa is supported by the H-bonding pattern observed in the wildtype NMR and simulated structures. In the NMR structure ensemble, half of the structures have no H1 interloop H-bonds, whereas our simulations show no intraloop H-bonds for the majority of the conformers. For 3-FFSa**,** there is also a well conserved interloop H-bond between Y32 and the residue at position 52a, which is observed during the data evaluation period not only in the type 1 H1 structures of this mutant, but also in those attained by the other mutants that display this canonical type (see Fig. 1.4). This specific interloop bond may be relevant for cooperative stabilization of H1 type 1 and H2 type 2A loops, given that it has been observed in some of the experimental structures that have these two canonical types (e.g., PDB structure 1TET).

Fig. 1.5 shows a visual comparison for the three loops of a representative 3-FFSa stable conformer found by simulation, the initial 3-FFSa conformation of the simulation, and the corresponding reference structures. The H1 loop side chains of F27, F29 and Y32 have rearranged to positions which are almost coincident with those of the equivalent 1DFB residues. While the side chains of F27 and F29 have been buried in the interior of the loop, that of Y32 remains at the surface of the domain with its aromatic ring slightly shielded by loops H1 and H3, and its hydroxyl group pointing outwards from the domain. No structures were found in which the Y32 side chain lies flat below the loop (as seen in the wildtype case), as the burial of side chains F27 and F29 precludes this from happening. This displacement of the Y32 side chain was also observed in all the low RMSD structures of other mutant cases, indicating

that the burial of any hydrophobic side chain at position 27 or 29 likely has the same effect on Y32.

Altogether, these observations corroborate that a stable type 1 H1 structure for the 3-FFSa mutant was obtained, with no noticeable perturbation of the H2 and H3 structures.

Considering the H1 loops of the remaining mutants, two particularly interesting cases are 1-Fb and 2-FS which, despite showing a reduced number of configurations with low H1 RMSD values from 1DFB, display a relatively stable behavior (see Table 1.3). A representative structure of the unconventional H1 loop displayed by these two mutants is shown in Fig. 1.6A. The F29 side chain of the representative 2-FS structure is driven away from W52a (not shown) and toward the canonical location of F27. A similar position for the F29 side chain is observed in the relatively large number of low H1 RMSD conformers of 3-FFSb (29%), but in this case the F27 side chain is forced out toward the solvent (Fig. 1.6B). Conformations with a non-buried F27 side chain are not likely to be energetically favorable since experimental structures show that this side chain is consistently buried in canonical H1 loops. The fact that 3-FFSb is found to be the most unstable mutant (see Table 1.3) supports the notion that its unfavorable F27 side chain positioning may lead to instability. Such mutants which cannot bury surface hydrophobic side chains may reduce VHH domain solubility (possibly promoting the aggregation of surface residues).

A plausible mechanism for the H1 non-canonical to canonical type 1 transition of the successful 3-FFSa was gathered from visual inspection of the trajectories of simulated conformations corresponding to 300 K (data not shown). Initially, two rapid and apparently cooperative changes take place: i) the Y32 side chain is driven out of the interior of the H1 loop and is exposed to the solvent; ii) the F29 side chain buries

within the loop at a position analogous to that observed for the low RMSD 1-Fb, 2-FS and 3-FFSb conformations (see 2-FS in Fig. 1.6A). These rapid transitions are later followed by two slow events: i) Burial of the F27 side chain at its canonical position, displacing F29 from the location observed for 2-FS in Fig. 1.6A to its canonical position (1DFB in Fig. 1.6A); ii) a shift of Y32 towards the H2 loop for subsequent formation of the stabilizing interloop H-bond with the residue at position 56.

The conformational behavior of H2 upon mutation is summarized in Tables 1.2 and 1.3. For the mutants studied, the type 2A conformation observed in the wildtype crystal structure is conserved. Upon inspection of the H2 loop of 3-FFSa (Fig. 1.3B), it is apparent that this loop closely maintains its initial type 2A structure throughout the entire run. Fig. 1.4 shows the characteristic H-bonding pattern of this mutant. The backbone-backbone H-bond between residues 52 and 56 is conserved in all H2 conformations examined. Table 1.3 shows average H2 RMSD values with respect to the typical structure ranging from 0.36-0.58 Å. A high degree of agreement between the simulated and reference structures of the H2 loop is also evident in Fig. 1.5. It is worth noting that none of the cases studied show any evidence of type 3 structures.

Regarding the H3 loop conformations, the RMSD values from 1HCV for the mutants studied are similar to those observed for the wildtype and the structures conserve a kinked shape. In general, both markers suggest that the influence of the mutations on the H3 loop structure is not significant. With respect to the H3 of 3-FFSa, a structure that closely resembles that of the simulated wildtype H3 loop (Fig. 1.2C) is observed (Fig. 1.5). Accordingly, these mutant H3 loops preserve their kinked conformation and have the well conserved pair of H-bonds between residues E96 and T99. However, it is noted that the H3 loop of 3-FFSb was observed to be somewhat more variable than that of the other mutants (see Table 1.3) and the wildtype.

Overall, the simulated mutations appear to significantly affect H1 loop stability, with average H1 RMSD values ranging from 0.51-1.29 Å from the typical structure. However, the overall stability of each mutant appears to be mainly, but not exclusively, due to the behavior of the H1 loop. Thus, the average variation of the three loops was used as an approximate measure of system stability. It can be seen in Table 1.3 that the highest variation is observed for 3-FFSb and 1-Fa, whereas 3-FFSa and 2-FS show the lowest deviation. The 2-FS H1 loop remains in a highly stable non-canonical conformation (see Fig. 1.6A), making it attractive for further study.

Candidate mutant screening via Multiple Mutant REM

It is known that the large number of degrees of freedom in an antigen binding site hampers accurate prediction of the key interactions that lead to the stabilization of a given target structure. In the absence of detailed knowledge of a reliable stabilization mechanism, site-directed mutagenesis methods like the one implemented in this study can, along with guidance from antibody databases, greatly aid the identification of candidate mutations. Furthermore, tools that facilitate the effective selection of potentially useful mutants for subsequent studies can substantially reduce the number of mutants to be analyzed. Thus, a quick screening method is very desirable. In this context, an alternative application of REM that leads to a fast evaluation of the relative stability among various mutants is described below.

In REM, the recursive exchanges of different configurations (replicas) of a given mutant drive the lower energy replicas to the lower temperature boxes. An analogous behavior can be expected from a single REM run in which different mutants are placed at one or more temperature boxes. In this way, the overall system minimizes its free energy by preferentially placing the lower energy mutants in the

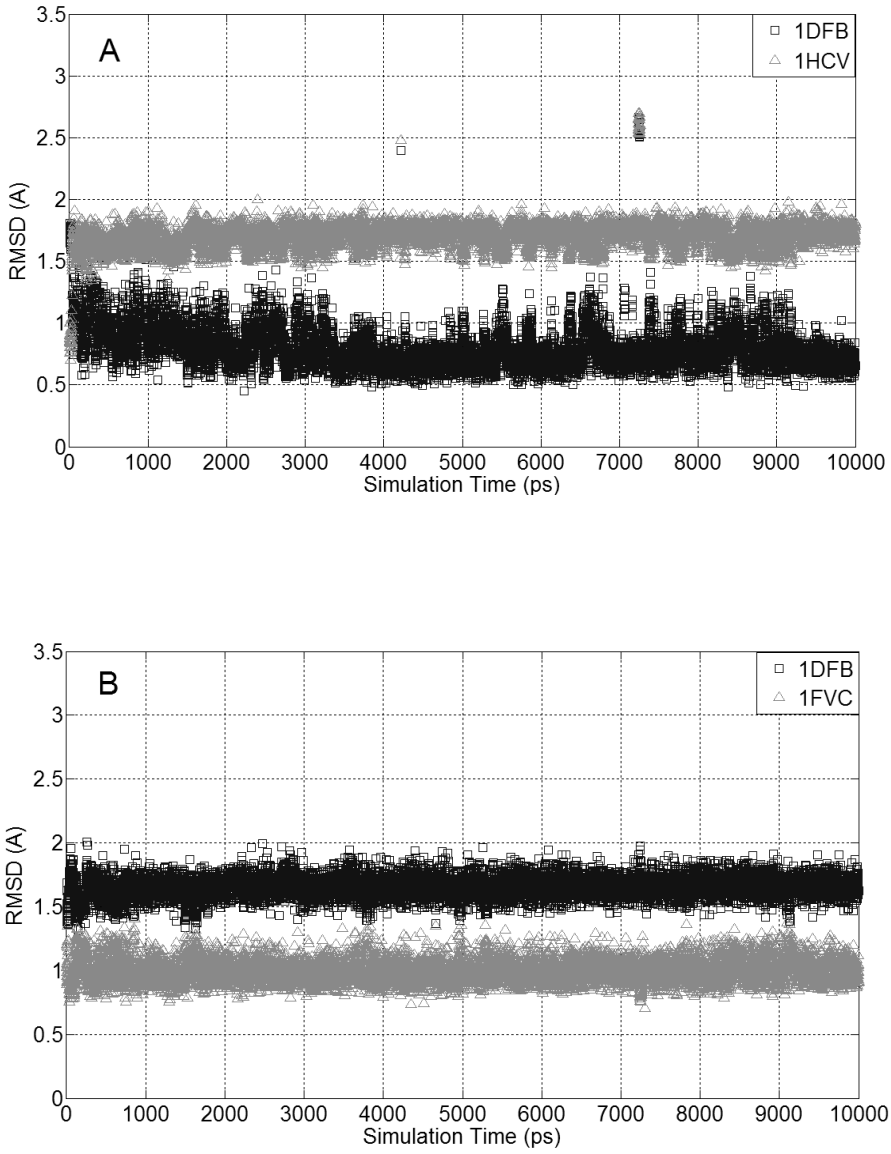


Figure 1.3. RMSD values between the conformations of the simulated H1 loop and 1DFB, 1HCV (*A*), and the simulated H2 loop and 1DFB, 1FVC (*B*), of a 10-ns run for the 3-FFSa case.

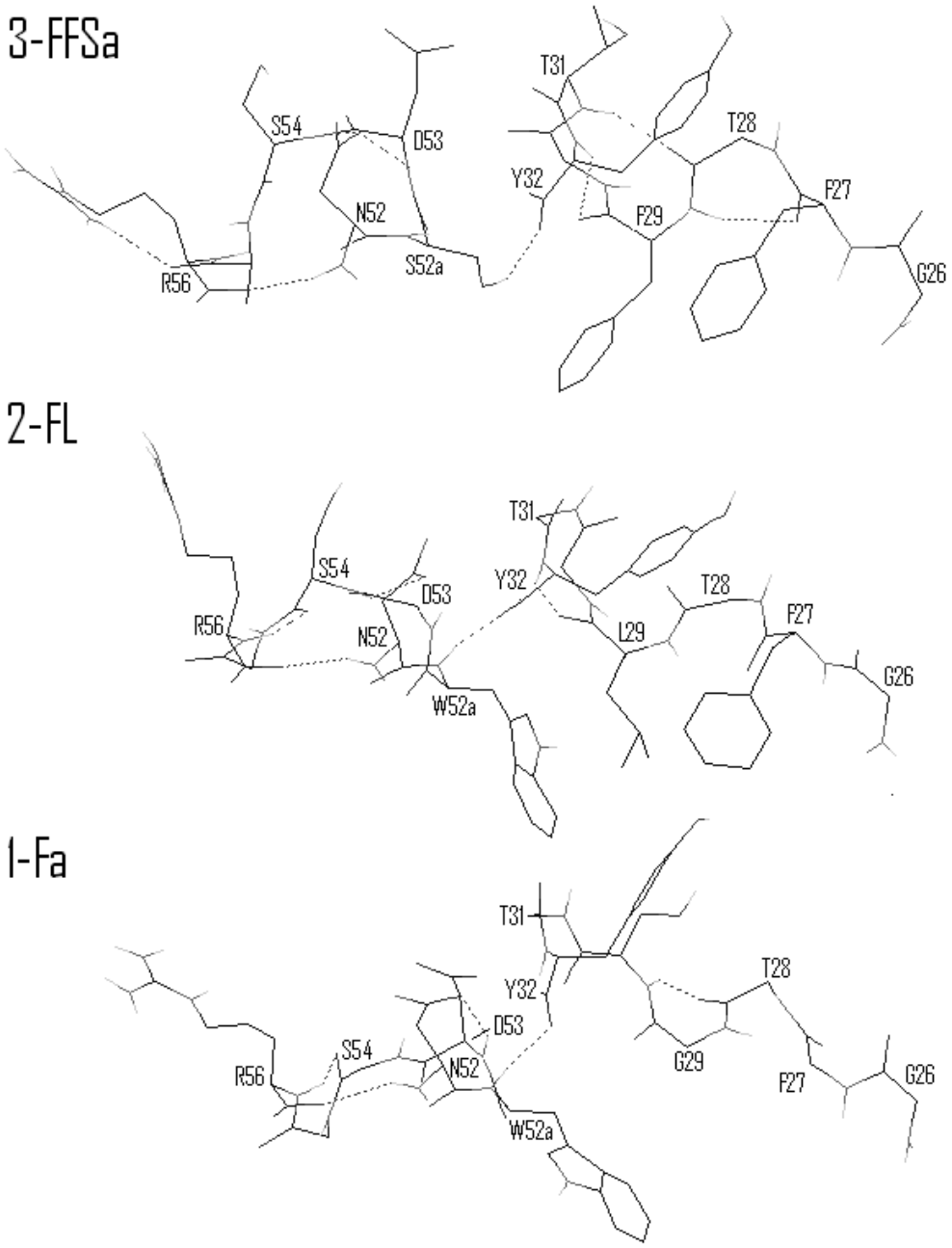


Figure 1.4. Simulated H1 (*residues 26-32*) and H2 (*residues 52,52a-56*) loops from the low RMSD structures of mutants 3-FFSa, 2-FL, and 1-Fa, showing the H-bonding pattern for each case. Only those side chains that are required to display the H-bonds are illustrated.

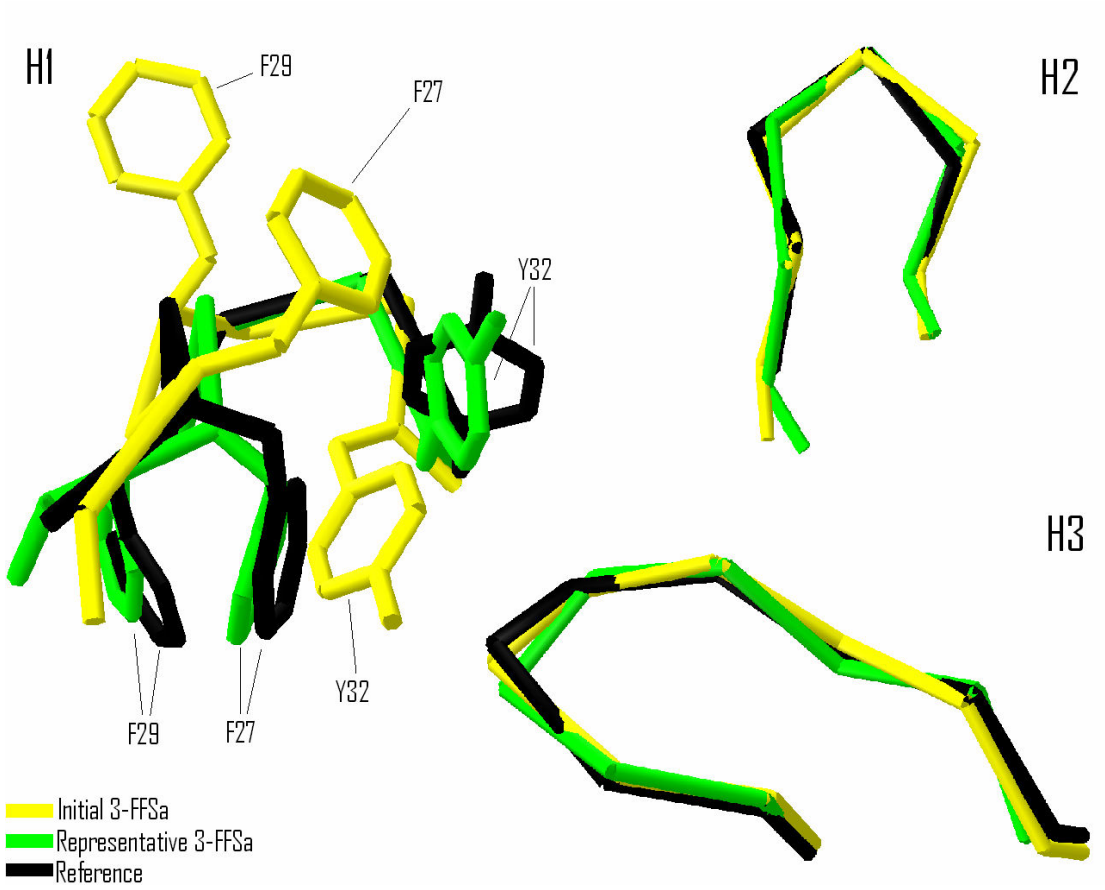


Figure 1.5. Simulated loops of a representative 3-FFSa structure, as compared to the initial 3-FFSa conformation used for the simulation and the reference structures 1DFB (*H1 loop*), 1FVC (*H2 loop*), and 1HCV (*H3 loop*). H1 side chains 27, 29 and 32 are also shown.

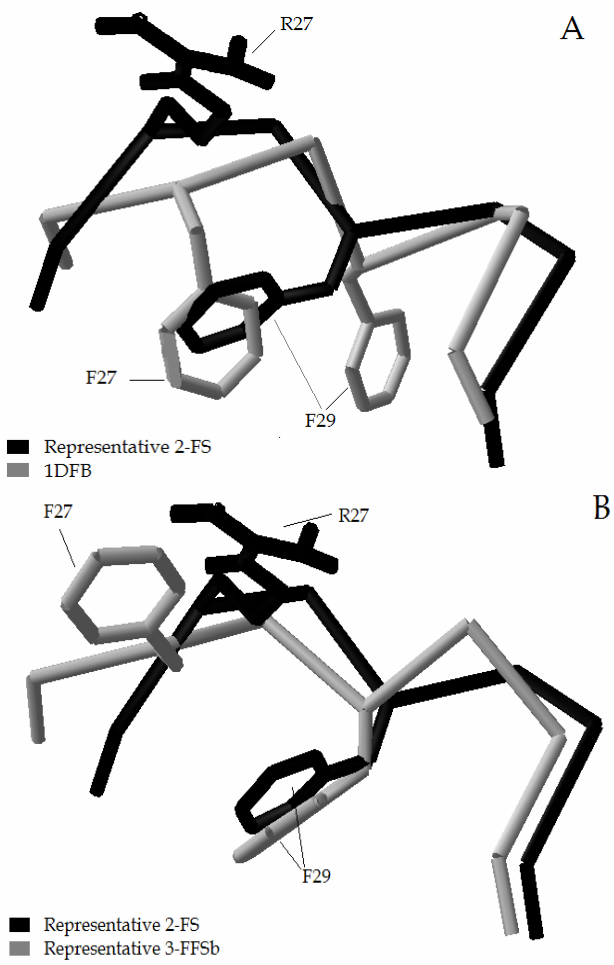


Figure 1.6. Representative structure of the simulated H1 loop of mutant 2-FS as compared to 1DFB H1 loop (*A*) and a low RMSD H1 loop of mutant 3-FFSb (*B*). Only relevant side chains are illustrated.

lower temperature boxes, and the higher energy mutants in the higher temperature boxes. The replica swaps should still enable enhanced configurational sampling to allow those candidates whose mutations can potentially drive the loops back to energetically favorable conformations, to achieve such low-energy stable states faster and to successfully compete for the low temperature boxes. The net outcome from the application of this method, to be denoted as multiple mutant REM or MMREM, is a quick sorting of the relative *stability* of the multiple mutants simulated. For the systems studied in this work, MMREM was validated by comparing the mean temperatures of residence of competing replicas with results for structural stability determined from REM simulations. Two variations of this approach, a rigorous and a simplified version, which can be readily implemented in various scenarios, are discussed in detail in Appendix B.

The simplified version of MMREM was first applied to a 12-replica system with six of the mutants given in Table 1.1, 1-Fb being excluded. Two replicas of each mutant were initially distributed throughout the 300-900°K range, as indicated in Table 1.4. Fig. 1.7 illustrates the resulting walk over temperature space for both replicas of two representative mutants, 3-FFSa and 3-FFSb, during a 10-ns run. As observed, the relative energies of the mutants encourage replicas to compete for temperature boxes, with the more stable ones (i.e., the ones that can more readily sample low energies) tending to remain at the lower temperatures. This trend is captured by the average temperatures visited by replicas during the data evaluation period (see Table 1.4). The results are in good agreement with findings from the conventional REM simulations. For example, the relative average loop variability (Table 1.3) coincides with the positioning of the mutants according to their most frequent temperature of residence using MMREM. Furthermore, the relative location of the mutants whose low H1 RMSD conformations displayed a type 1 structure upon

visual inspection, concurs with their frequency of adopting a type 1 H1 structure in REM (3-FFSa, 2-FF, 2-FL and 1-Fa, in this order).

To further assess the MMREM approach, a system with 12 replicas each with a different mutant, was simulated for a period of 10 ns. In addition to the cases used in the 6-mutant run described above, the mutants considered were: 1-Fb, 2-FI (R27F,G29I), 3-FFD (R27F,G29F,W52aD), 3-FFP (R27F,G29F,W52aP), 3-FLS (R27F,G29L,W52aS) and 4-FFSS (R27F, G29F, W52aS, N52S). Table 1.5 shows the initial temperatures and most visited temperature for each mutant**.** As before, 3-FFSb and 1-Fa locate themselves at the highest temperatures, now joined by 4-FFSS. The last mutant is unable to achieve low energy states, despite having the three stabilizing mutations R27F, G29F and W52aS. On the other hand, the most stable cases found via conventional REM, namely 3-FFSa and 2-FS, have the third and fourth lowest temperatures of residence, respectively, preceded by mutants 3-FFD and 1-Fb. The latter case showed high stability during the REM runs, while the former is detailed below. Significantly, the two groups of mutants that are believed to have a similar behavior given the nature and position of their mutations, 3-FFSa/3-FLS and 2-FF/2-FI/2-FL, locate themselves at contiguous boxes.

A novel finding from the 12-MMREM simulation is that the replicas from the newly introduced 3-FFD mutant are found to maintain the lowest energies. Thus, in an attempt to validate the MMREM result, a 10-ns conventional REM run was conducted for this mutant. Fig. 1.8 illustrates the RMSD values for the H1 and H2 loops, with respect to their corresponding reference structures. The 3-FFD mutant H1 loop (Fig. 1.8A) is seen to reach a stable H1 type 1 structure even faster than the successful 3-FFSa case (see Fig. 1.3A), with 99% of the configurations having a RMSD value below 1.2 Å with respect to 1DFB and an average RMSD of the three loops from the typical structure of 0.49 Å (see Table 1.3 for comparison) for the data evaluation

period. Fig. 1.9 shows a visual inspection of the H1 and H2 loops from a representative 3-FFD structure with low RMSD values, highlighting their corresponding interactions. The structure, side chain location and H-bonding pattern of loops H1 and H2 closely resemble those observed for low energy 3-FFSa structures (see Fig. 1.4). In addition, 98% of the H2 loop structures have RMSD values below 1.2 Å with respect to 1FVC (Fig. 1.8B), conformations that are stabilized by three intraloop backbone-backbone H-bonds. As in the 3-FFSa case, the 3-FFD H3 structures resemble those observed for the wildtype H3 loop. The increased stability of the 3-FFD mutant may be due to a slight structural change in the orientation of D53 encouraged by the extension of the D52a side chain toward the solvent, promoting a strong interaction between D53 and S30 that keeps the H1 loop conformations in place.

It is pointed out that the MMREM results reported here should be interpreted with care given that: (i) the simulation periods were relatively short, (ii) only the simplified version of the method was implemented, and (iii) the simulation parameters were not optimized. For example, there is likely an optimal ratio of replicas to mutants which leads to faster equilibration (too many mutants may hamper a broad enough exploration of the temperature space for a given mutant to achieve ergodic sampling). The simplified MMREM is also likely to be most effective when mutants differ only by a few point mutations and when the system is sufficiently constrained so that a unique structure can be associated with the most energetically favorable conformation achievable. In this way, all mutants could be seen as competing for the same structure and where the "winner" is the one that best stabilizes that structure at lower temperatures. The method could be further refined to make it more specific in targeting a desired structure, for example, by introducing in addition to temperature, another "tempering" REM parameter that can more directly capture deviations from

Table 1.4. Initial temperature distribution and average between the most visited temperatures for the two replicas of each mutant, for the data evaluation period of a 10-ns 6-mutant MMREM run.

Case	Initial Ts for each replica ($^{\circ}K$)	Average between most visited Ts ($^{\circ}K$)
1-Fa	300, 546	702
2-FF	366, 667	468
2-FS	447, 737	450
2-FL	332, 604	497
3-FFSa	405, 815	316
3-FFSb	494, 900	858

Table 1.5. Initial temperature distribution and most visited temperature for each mutant, for the data evaluation period of a 10-ns 12-mutant MMREM run.

Case	Initial Ts for each replica ($^{\circ}K$)	Most visited T ($^{\circ}K$)
1-Fa	300	815
1-Fb	332	332
2-FF	366	604
2-FS	405	494
2-FL	447	546
2-FI	494	546
3-FFSa	546	366
3-FFSb	604	900
3-FFP	667	667
3-FFD	737	300
3-FLS	815	405
4-FFSS	900	737

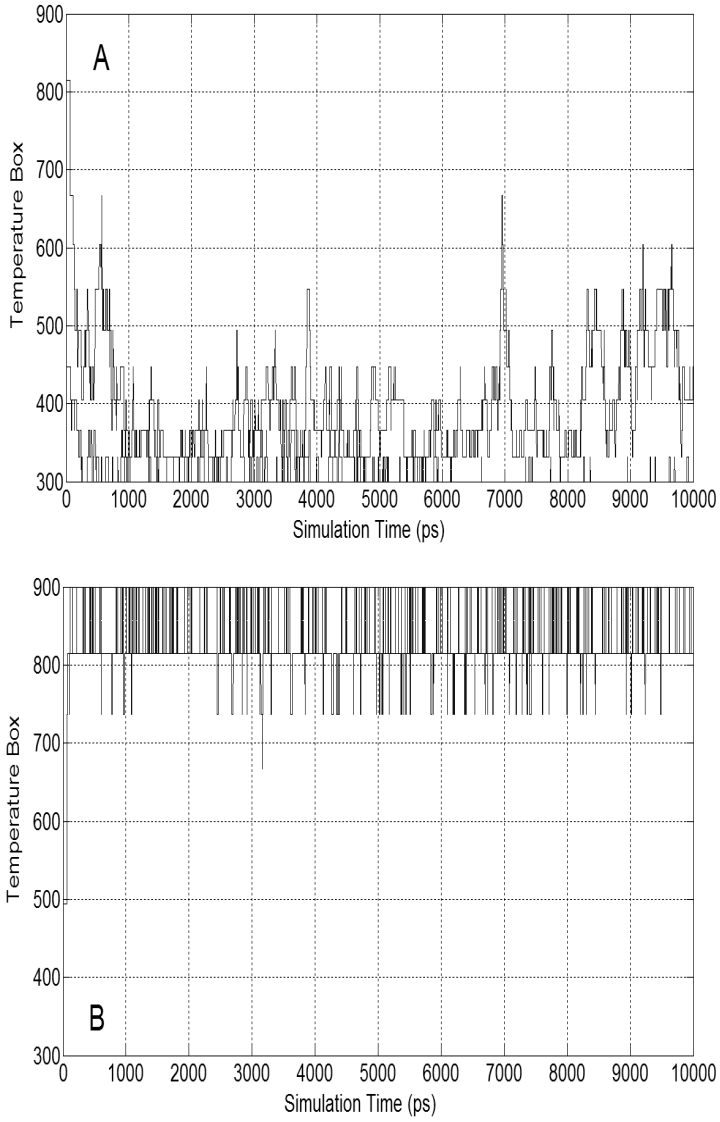


Figure 1.7. Walk over temperature space for the data evaluation period of a 6-mutant MMREM run. The two replicas for mutants 3-FFSa (*A*) and 3-FFSb (*B*) are displayed in each plot.

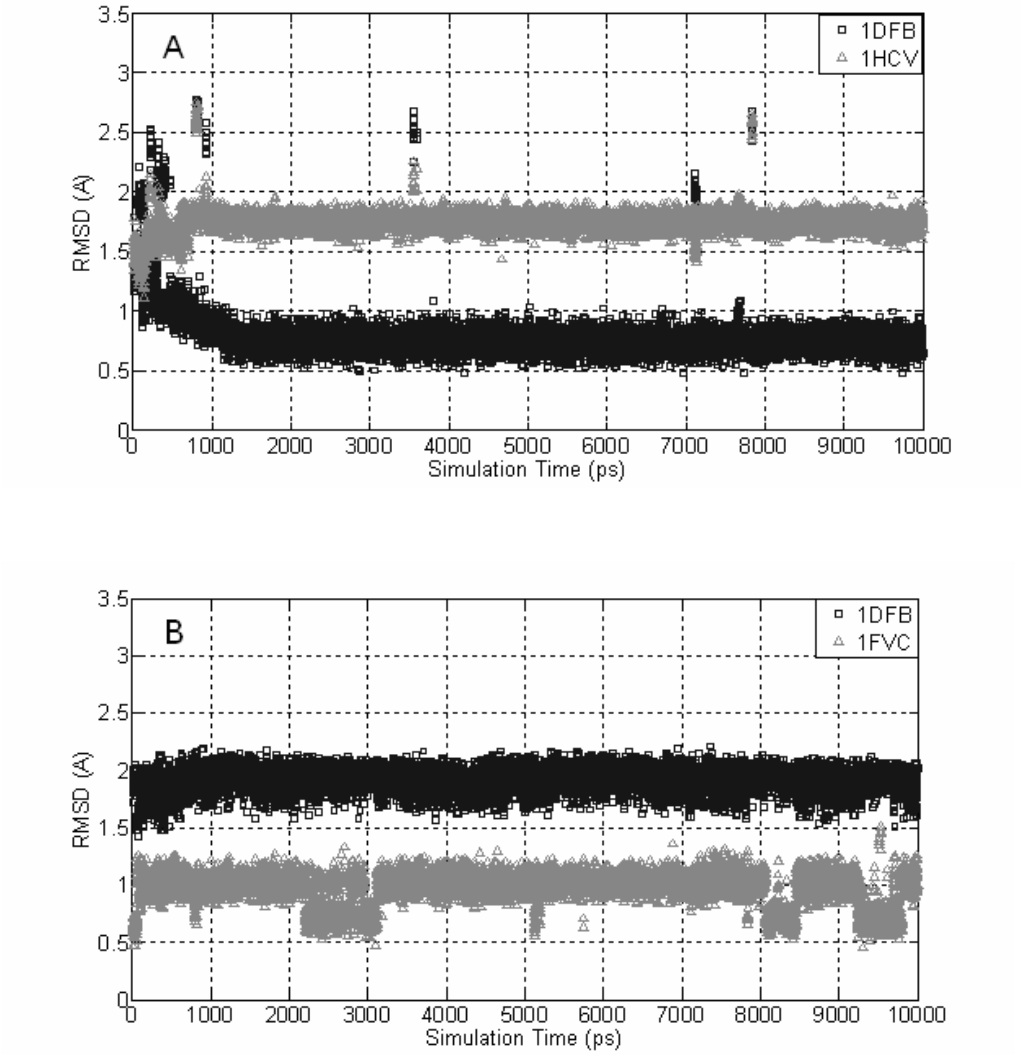


Figure 1.8. RMSD values between the conformations of the simulated H1 loop and 1DFB, 1HCV (*A*), and the simulated H2 loop and 1DFB, 1FVC (*B*), of a 10-ns run for the 3-FFD case.

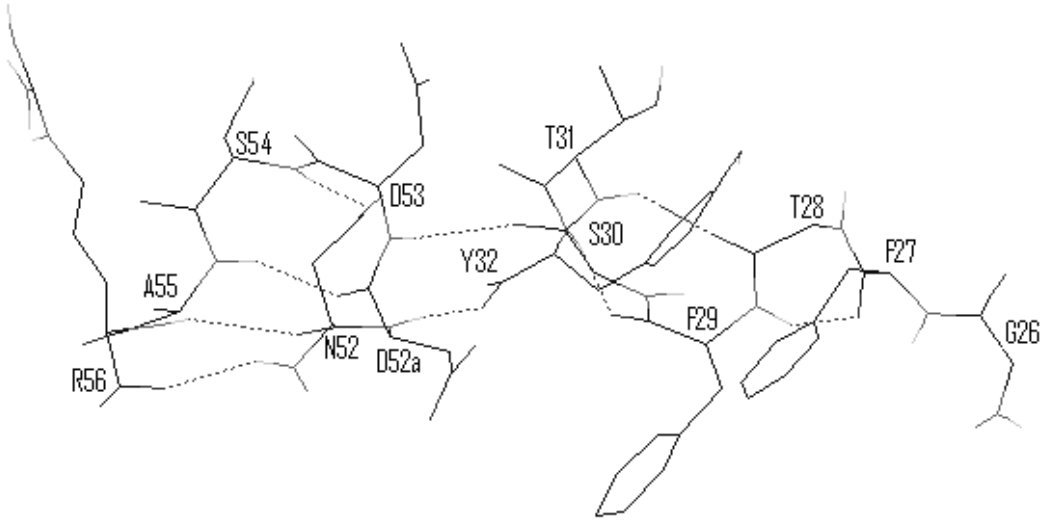


Figure 1.9. Simulated H1 (*residues 26-32*) and H2 (*residues 52,52a-56*) loops of a representative 3-FFD structure, showing its H-bonding pattern.

the target structure (like the RMSD values of the positions or angles of key residue atoms).

VI. FINAL REMARKS

The set of VHH antibody systems simulated in the present study constitute a challenging application for available simulation protocols and force fields. The three loops of the wildtype and mutant systems display considerable variation in hydrophilicity and structural flexibility. Accordingly, we found that adequate study of their conformational equilibria demands the use of advanced methods and a careful examination of force fields. For the applications presented, we observed that REM in combination with the CHARMM19-GBMVA force field is capable of generating realistic structural ensembles from 10 ns simulations.

One of our primary goals was to use this combination of methods and force field to identify specific residues that reshape the H1 loop of the llama VHH domain 1HCV into a stable type 1 canonical conformation. Furthermore, by studying the structural changes of the three hypervariable loops of the wildtype and mutant structures, insights were obtained into the possible reasons underlying the conformational diversity of each particular system. A novel formulation of REM was also developed for screening multiple mutants based on their relative proclivity to adopt energetically favorable structures with associated reduced configurational dispersion. In general, the study of loop dynamics by simulation significantly aids in their structural typification and complements the information obtained from NMR and crystal analyses. In addition, the systematic approach implemented here may be useful in antibody engineering applications that require grafting loops having specific functionalities onto the framework regions of VHH domains.

The type 1 canonical structure of H1 is vastly conserved in human and mouse antibodies, implying that its backbone structural invariance has functional significance. In this work, it was observed that a type 1 H1 structure has increased stability over other H1 loop conformations, consistent with its high occurrence in databases of antibody crystal structures. Two cases with a highly stable type-1 H1 loop structure were obtained from the mutant simulations, each having three point mutations, namely 3-FFSa and 3-FFD. These results confirm the marked influence of highly hydrophobic residues (e.g., phenylalanine) at positions 27 and 29, whose stable side chain positioning drives the H1 backbone to a type 1 structure with a conserved H-bonding pattern. This appears to occur, however, only if clashing of residue W52a with residue 29 is avoided (in these two cases, the smaller residues S and D prevent clashing with F29). Mutants with one (1-Fa) and two (2-FF and 2-FL) mutations were also able to achieve the target H1 conformation, but with a moderate to low occurrence; such cases may be useful for further analysis in applications in which an increased number of mutations in a therapeutic antibody is detrimental to its immunocompatibility. It is worth noting that the achievement of an H1 type 1 conformation was observed to be dependent upon mutations in both the H1 and H2 loops, which indicates that interloop interactions may be relevant for the prediction of canonical conformations, at least for particular cases such as the ones studied in this work.

Overall, the average system stabilities measured in REM simulations are consistent with the average temperatures sampled by replicas in MMREM simulations. Furthermore, cases with successful mutations (i.e., that achieve high conversion to a type 1 H1 structure) have increased loop stability. Conversely, some of the tested mutants that show decreased stability display an H1 loop variability comparable to that of the wildtype fragment. Interestingly, two highly stable cases with a coincident

non-canonical H1 loop were found; unique structures such as these may contribute to the design of novel binding domains.

In addition to the large binding repertoire observed for VHHs in camelids due to sequence variability**,** each loop possesses internal motion that promotes the sampling of distinct backbone conformations in some antibodies. In particular, our simulations of the llama anti-hCG VHH fragment suggest that its H1 loop has considerable backbone flexibility, in agreement with results of other studies.[22,43,44] Such flexibility presumably has functional significance, perhaps promoting induced-fit binding to hCG that results in a moderate affinity of K_d=300 nM.[44]

In general, a lock and key recognition mechanism will be promoted by stable loops, leading to improved affinity for some antigens.[64] Given that the MMREM method is oriented towards evaluating the relative stabilities of different mutants, it may play an important role within a rational antibody design protocol aimed at engineering loops that bind antigens based on such a mechanism. Through the identification of stable hypervariable loops, the method may also be useful in therapeutic applications[65,66] for reducing antibody immunogenicity of otherwise variable non-canonical loops that may be able to bind an increased number of targets. In this way, experimental procedures such as VHH loop grafting of highly stable structures could be more effectively realized.

Ongoing efforts are focused on increasing the efficiency of REM and MMREM simulations toward explicit-solvent mutagenesis analyses of complex structures to more accurately describe water mediated interactions (known to be of crucial importance in some systems).

ACKNOWLEDGEMENTS

The authors acknowledge the financial support of the National Science Foundation, awards BES-0093769 and ECS-0304483, and the US Department of Energy, Grant No. DE-FG02-05ER15682.

REFERENCES

1. B. Al-Lazikani, A. M. Lesk, and C. Chothia, J. Mol. Biol. **273**, 927 (1997).

2. C. Chothia and A. M. Lesk, J. Mol. Biol. **196**, 901 (1987).

3. C. Chothia, A. M. Lesk, E. Gherardi, I. M. Tomlinson, G. Walter, J.D. Marks, M.B. Llewelyn, and G. Winter, J. Mol. Biol., **227**, 799 (1992**)**

4. C. Chothia, A. M. Lesk, A. Tramontano, M. Levitt, S. J. Smithgill, G. Air, S. Sheriff, E. A. Padlan, D. Davies, W. R. Tulip, P. M. Colman, S. Spinelli, P. M. Alzari, , and R. J. Poljak, Nature **342**, 877 (1989**)**.

5. K. Decanniere, S. Muyldermans, and L. Wyns, J. Mol. Biol. **300**, 83 (2000).

6. P. Delapaz, B. J. Sutton, M. J. Darsley, and A. R. Rees, EMBO J. **5,** 415 (1986).

7. S. Ewert, A. Honegger, and A. Pluckthun, Methods **34**, 184 (2004).

8. G. Johnson, and T. T. Wu, Int. Immunol. **10**, 1801 (1998).

9. C. Mandal, B. D. Kingery, J. M. Anchin, S. Subramaniam, and D. S. Linthicum, Nat. Biotechnol. **14**, 323 (1996).

10. A. C. R. Martin, J. C. Cheetham, and A. R. Rees, Methods Enzymol. **203**, 121 (1991).

11. A. C. R. Martin and J. M. Thornton, J. Mol. Biol. **263**, 800 (1996).

12. E. Monsellier and H. Bedouelle, J. Mol. Biol. **362**, 580 (2006).

13. V. Morea, A. M. Lesk, and A. Tramontano, Methods **20**, 267 (2000).

14. V. Morea, A. Tramontano, M. Rustici, C. Chothia, and A. M. Lesk, J. Mol. Biol. **275**, 269 (1998).

15. P. A. Ramsland, L. W. Guddat, A. B. Edmundson, and R. L. Raison, J. Comput. Aided Mol. Des. **11**, 453 (1997).

16. M. Reczko, A. C. R. Martin, H. Bohr, and S. Suhai, Protein Eng. **8**, 389 (1995).

17. A. Schlessinger, Y. Ofran, G. Yachdav, and B. Rost, Nucleic Acids Res. **34**, D777 (2006).

18. H. Shirai, A. Kidera, and N. Nakamura, FEBS Lett. **455**, 188 (1999).

19. S. J. Smithgill, C. Mainhart, T. B. Lavoie, R. J. Feldmann, W. Drohan, and B. R. Brooks, J. Mol. Biol. **194**, 713 (1987).

20. R. E. Bruccoleri, E. Haber, and J. Novotny, Nature **335**, 564 (1988).

21. A. K. Felts, E. Gallicchio, D. Chekmarev, K. A. Paris, R. A. Friesner, and R. M. Levy, Journal of Chemical Theory and Computation **4**, 855 (2008).

22. M. K. Fenwick, and F. A. Escobedo, Biopolymers **68**, 160 (2003).

23. R. M. Fine, H. Wang, P.S. Shenkin, D.L. Yarmush, and C. Levinthal, Proteins **1,** 342 (1986).

24. J. I. Higo, V. Collura, and J. Garnier, Biopolymers **32**, 33 (1992).

25. V. Hornak and C. Simmerling, Proteins Struct. Funct. Genet. **51**, 577 (2003).

26. S. T. Kim, H. Shirai, N. Nakajima, J. Higo, and Nakamura, H., Proteins Struct. Funct. Genet. **37**, 683 (1999).

27. M. Krol, J. Comput. Chem. **24**, 531 (2003).

28. J. L. Pellequer and S. W. Chen, Biophys. J.**73**, 2359 (1997).

29. P. S. Shenkin, D. L. Yarmush, R. M. Fine, H. J. Wang, and C. Levinthal, Biopolymers **26**, 2053 (1987).

30. N. Sinha and S. J. Smith-Gill, Cell Biochem. Biophys. **43**, 253 (2005).

31. C. Tenette, F. Ducancel, and J. C. Smith, Proteins Struct. Funct. Genet. **26**, 9 (1996).

32. S. Voordijk, T. Hansson, D. Hilvert, and W. F. van Gunsteren, J. Mol. Biol. **300**, 963 (2000).

33. Q. Zheng, R. Rosenfeld, C. Delisi, and D. J. Kyle, Protein Sci. **3**, 493 (1994).

34. A. C. R. Martin, J. C. Cheetham, and A. R. Rees, Proc. Natl. Acad. Sci. U. S. A. **86**, 9268 (1989).

35. M. T. Mas, K. C. Smith, D. L. Yarmush, K. Aisaka, and R. M. Fine, Proteins Struct. Funct. Genet. **14**, 483 (1992).

36. D. M. Webster, S. Roberts, J. C. Cheetham, R. Griest, and A. R. Rees, Int. J. Cancer **Suppl. 3**, 13 (1988).

37. A. Fiser, R. K. G. Do, and A. Sali, Protein Sci. **9**, 1753 (2000).

38. C. Hamers-Casterman, T. Atarhouch, S. Muyldermans, G. Robinson, C. Hamers, E. B. Songa, N. Bendahman, and R. Hamers, Nature **363**, 446 (1993).

39. R. H. J. van der Linden, L. G. J. Frenken, B. de Geus, M. M. Harmsen, R. C. Ruuls, W. Stok, L. de Ron, S. Wilson, P. Davis, and C. T. Verrips, BBA- Protein Struct. M. **1431**, 37 (1999).

40. J. M. J. Perez, J. G. Renisio, J. J. Prompers, C. J. van Platerink, C. Cambillau, H. Darbon, and L. G. Frenken, Biochemistry (Mosc), **40**, 74 (2001).

41. M. Lauwereys, M. A. Ghahroudi, A. Desmyter, J. Kinne, W. Holzer, E. De Genst, L. Wyns, and S. Muyldermans, EMBO J. **17**, 3512 (1998).

42. D. Saerens, M. Pellis, R. Loris, E. Pardon, M. Dumoulin, A. Matagne, L. Wyns, S. Muyldermans, and K. Conrath, J. Mol. Biol. **352**, 597 (2005).

43. J. G. Renisio, J. Perez, M. Czisch, M. Guenneugues, O. Bornet, L. Frenken, C. Cambillau, and H. Darbon, Proteins Struct. Funct. Genet. **47**, 546 (2002).

44. S. Spinelli, L. Frenken, D. Bourgeois, L. deRon, W. Bos, T. Verrips, C. Anguille, C. Cambillau, and M. Tegoni, Nat. Struct. Biol. **3**, 752 (1996).

45. P. Brenner, C. R. Sweet, D. VonHandorf, and J. A. Izaguirre, J. Chem. Phys. **126**, - (2007).

46. Y. Sugita, and Y. Okamoto, Chem. Phys. Lett. **314**, 141 (1999).

47. B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus, J. Comput. Chem. **4**, 187 (1983).

48. M. Feig, J. Karanicolas, and C.L. Brooks III, MMTSB Tool Set, MMTSB NIH Research Resource, The Scripps Research Institute (2001).

49. X. M. He, F. Ruker, E. Casale, and D. C. Carter, Proc. Natl. Acad. Sci. U. S. A. **89**, 7154 (1992).

50. C. Eigenbrot, M. Randal, L. Presta, P. Carter, and A. A. Kossiakoff, J. Mol. Biol. **229**, 969 (1993).

51. V. K. Nguyen, R. Hamers, L. Wyns, and S. Muyldermans, EMBO J. **19**, 921 (2000).

52. I. M. Tomlinson, G. Walter, J. D. Marks, M. B. Llewelyn, and G. Winter, J. Mol. Biol. **227**, 776 (1992).

53. J. P. Ryckaert, G. Ciccotti, and H. J. C. Berendsen, J. Comput. Phys. **23**, 327 (1977).

54. D. J. Earl and M.W. Deem, PCCP **7**, 3910 (2005).

55. A. D. Mackerell, M. Feig, and C. L. Brooks III, J. Comput. Chem. **25**, 1400 (2004).

56. W. P. Im, M. S. Lee, and C. L. Brooks III, J. Comput. Chem. **24**, 1691 (2003).

57. E. Neria, S. Fischer, and M. Karplus, J. Chem. Phys. **105**, 1902 (1996).

58. M. S. Lee, F. R. Salsbury, and C. L. Brooks III, J. Chem. Phys. **116**, 10606 (2002).

59. T. Lazaridis and M. Karplus, Proteins Struct. Funct. Genet. **35**, 133 (1999).

60. M. A. Olson, M. Feig, and C. L. Brooks III, J. Comput. Chem., Early View (2007).

61. A. Kone, and D. A. Kofke, J. Chem. Phys. **122**, - (2005)

62. Q. L. Yan and J. J. de Pablo, J. Chem. Phys. **111**, 9509 (1999).

63. S. Trebst, M.Troyer, and U. H. E. Hansmann, J. Chem. Phys. **124**, 174903 (2006).

64. G. J. Wedemayer, P. A. Patten, L. H. Wang, P.G. Schultz, and R.C. Stevens, Science **276**, 1665 (1997).

65. L. R. Helms, and R. Wetzel, Protein Sci. **4**, 2073 (1995).

66. N. Lonberg, Nat. Biotechnol. **23**, 1117 (2005).

67. C. H. Bennett, J. Comput. Phys. **22**, 245 (1976).

68. A. A. Barker, Aust. J. Phys. **18**, 119 (1965).

CHAPTER 2

KINETICS AND REACTION COORDINATE FOR THE ISOMERIZATION OF ALANINE DIPEPTIDE BY A FORWARD FLUX SAMPLING PROTOCOL*

I. INTRODUCTION

Path sampling schemes such as forward flux sampling (FFS)[1-3] allow the computation of rate constants by overcoming the problems associated with simulating rare events (i.e., reducing the CPU time wasted on the uneventful waiting time between events). In FFS, interfaces are used to partition the phase space along an order parameter λ connecting the initial and final regions of interest; transition pathways between such regions are constructed by stitching together partial paths generated between successive interfaces. FFS has been successfully applied to Monte Carlo (MC) simulations[2-6] of complex systems involving rare events. However, it is of interest to use FFS with trajectories generated by molecular dynamics (MD), given the sampling efficiency observed in numerous MD studies for a broad range of systems. Furthermore, MD is particularly convenient for use with elaborate force fields in continuum simulations, which are prevalent in the study of proteins and other biomolecular systems. Although alternate path sampling schemes that use MD have been successfully applied for kinetic studies of rare events in biological systems (for a comprehensive review of path sampling-based studies see Dellago and Bolhuis),[7,8] FFS exhibits a number of strengths (such as simplicity and the ability to describe non-equilibrium systems) that make it an appealing choice for such studies.

---

* Reproduced with permission from C. Velez-Vega, E.E. Borrero, and F. A. Escobedo, J. Chem. Phys. **130**, 225101 (2009).

To the best of our knowledge, FFS with MD has only been used once very recently for the folding of a small protein) where it was found to lead to a transition path ensemble and a rate constant in gross error because the sampled trajectories were correlated around non-representative transition paths.[9] For our applications, such sampling problems are avoided by optimizing the order parameter λ and the positions of the interfaces along such a parameter. In particular, the suitable positioning of the first interface λ_0 is crucial: if λ_0 is too close to the initial basin then the crossing points (which serve as starting points of all trajectories) are abundant but very correlated, while if λ_0 is too far, then crossing points are well uncorrelated but too costly to generate. Accordingly, a method is introduced to find the optimal positioning of λ_0 by minimizing the CPU time needed to produce an uncorrelated crossing point.

In this work, the kinetics of alanine dipeptide in vacuum and in explicit solvent was studied via FFS-MC and FFS-MD simulations. Alanine dipeptide has been widely used for computational studies given its small size and its ability to display transitions between some of its preferred conformers in short timescales. Moreover, despite its simplicity the molecule is able to adopt, in aqueous environment, all conformations observed for α helix and β strand motifs in proteins.[10] Fig. 2.1 illustrates the four main torsion angles of alanine dipeptide (θ,ϕ,ψ,ζ). Although the thermodynamics of this system has been well characterized by experimental and theoretical methods,[11-14] the conformational kinetics (i.e., transition rate constant and reaction coordinate) of this molecule is still an active area of research. The ψ and ϕ dihedral angles are commonly considered to be good indicators of the conformational diversity of this peptide.[15-17] In vacuum, the free energy landscape obtained using CHARMM all-atom force field[18] and projected using these angles (ψ and ϕ) shows two distinct stable basins corresponding to states $C7_{eq}$ and C5. Although

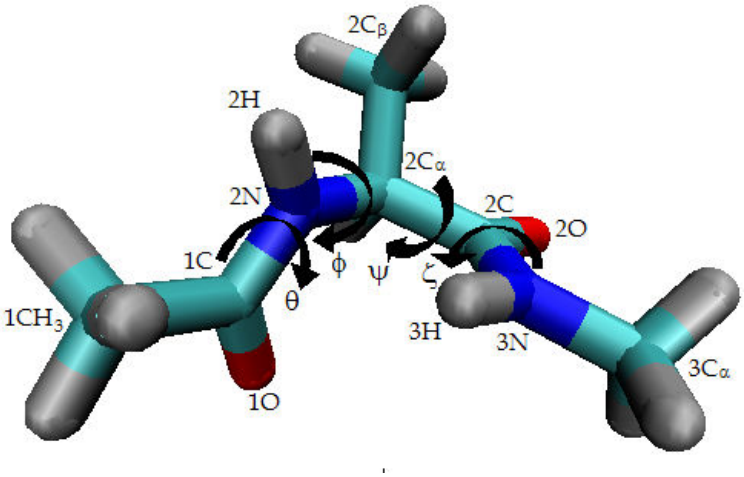


Figure 2.1. A model for alanine dipeptide. Also shown are the main dihedral angles: θ (O-C-N-C_α), ϕ (C-N-C_α-C), ψ (N-C_α-C-N), and ζ (C_α-C-N-H). Carbon, oxygen, nitrogen and hydrogen atoms are depicted in light green, red, blue and gray color, respectively.

these order parameters satisfactorily describe the system's distinct stable states, this does not imply that they will provide an accurate description for the dynamics of the transition. Thus, other variables (in addition to ψ and ϕ) and/or interaction terms between variables may also be important in the reaction coordinate model.

In contrast, the free energy landscape projected using ψ and ϕ dihedrals for alanine dipeptide in explicit solvent shows several minima. In this work, we focus on the $\beta_2/\alpha_R \Leftrightarrow C5/C7_{eq}$ transition. Various researchers have estimated transition rate constant values for the forward[10,17,19] and reverse transitions,[10,15] as well as the collective variables that are important for the description of the $\beta_2/\alpha_R \Leftrightarrow C5/C7_{eq}$ transitions.[15-17] For example, Bolhuis et al.[15] found that the solvent degrees of freedom may play a role in this transition and suggested their incorporation in the reaction coordinate model of the process. More recently, Ma and Dinner[16] performed an exhaustive search of many possible reaction coordinates for the forward transition using a genetic neural network (GNN) approach and concluded that the best reaction coordinate model for an adequate description of the alanine dipeptide isomerization should contain a term involving the torque around a specific bond, associated with electrostatic forces exerted by the water molecules in a particular hydrogen atom of the peptide. Hence, the second aim of our work is to use the FFS-LSE approach to obtain good reaction coordinate models for the $\beta_2/\alpha_R \Leftrightarrow C5/C7_{eq}$ transitions in explicit solvent and compare them with the aforementioned studies. Our strategy is thus to take full advantage of the adaptive algorithm[6] to obtain an optimized λ phase staging (that reduces the statistical error in the rate constant estimation) and set up the staging for FFS-LSE[5] simulations, to subsequently obtain a good estimate for the reaction coordinate.

By way of background, we start by briefly reviewing the FFS-type simulation scheme for the calculation of rate constants and transition pathways (Sec. II A), the

FFS-LSE algorithm (Sec. II B), and the adaptive algorithm which optimizes the phase space sampling (Sec. II C). The details for the stochastic approach employed in the FFS-MD simulations are given in Sec. II D. In Secs. III A and B, we give the simulation details for the alanine dipeptide system both in vacuum and in explicit solvent, respectively. In Sec. IV, we report values for the transition rate constant and estimates for the best reaction coordinate model. In Sec. V, we provide some concluding remarks.

II. METHODS

A. Forward Flux Sampling (FFS)

In this work, we used the Branched Growth method (BG) sampling scheme to generate transition paths [i.e., the transition path ensemble (TPE)] by a FFS-type approach.[1-3] The BG method is illustrated schematically in Fig. 2.2, where branched transition paths are generated one by one. The phase space is partitioned by employing a series of nonintersecting interfaces (n+1) such that the system is considered to be in region A for $\lambda(x) \leq \lambda_A(x)$ and in region B for $\lambda(x) \geq \lambda_B(x)$. The interfaces are defined by an order parameter $\lambda(x)$ (where x represents the phase space coordinates) whose value increases monotonically as the interfaces come close to region B. The TPE is generated in such a way that any trajectory from A to B passes through each interface and all the transitions between interfaces are free to follow any possible path between A and B.

The rate constant $k_{\mathrm{A}\to\mathrm{B}}$ of the process is defined as an average rate of transitions from two well-defined states A and B using an "effective positive flux" expression:[2,3,20,21]

$$k_{A\to B} = \overline{\Phi}_{A,0} P(\lambda_{n=B} \mid \lambda_0) \tag{1}$$

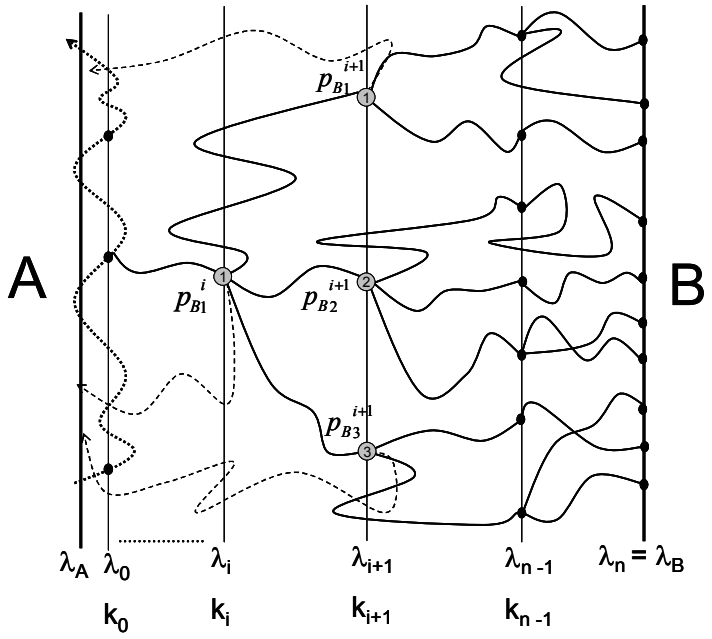


Figure 2.2. A schematic view of the generation of branched paths (thick lines) using the branched growth (BG) sampling method. The first stage involves the simulation run in the A basin shown by a dotted line. Starting points for the subsequent generation of branched paths are marked with a black circle at λ_0. The second stage corresponds to the trial runs (k_i) fired from λ_i ; those that reached the next λ_{i+1} interface are shown by a thick line and those which failed to reach λ_{i+1} are shown by a dotted line. For example, the $p_{B1}^{\ i}$ value for the point 1 at λ_i is then obtained recursively from Eq. (3): $p_{B1}^{\ i} = \left[p_{B1}^{\ i+1} + p_{B2}^{\ i+1} + p_{B3}^{\ i+1}\right]/4 = \left[1/2 + 2/3 + 1/2\right]/4$.

where $\overline{\Phi}_{A,0}$ is the total average flux of trajectories from A to λ_0, and $P(\lambda_{n=B} \mid \lambda_0)$ is the probability that a trajectory reaching λ_0 from A will reach B without returning to A.[1] In the first stage of the algorithm, the flux term, $\overline{\Phi}_{A,0}$, is calculated by carrying out a simulation in the basin of attraction of A, where $\lambda(x)$ is monitored and configurations crossing λ_0 are stored. In the second stage of the algorithm, a branched path is generated from a randomly chosen configuration at λ_0 and an estimate for $P(\lambda_{n=B} \mid \lambda_0)$ value is obtained by initiating k_0 trial runs which are continued until either reaching λ_1 or returning to the initial region. For each configuration at λ_1 (i.e., N_S^1), k_1 trial runs are then started to λ_2 or back to A. This procedure is repeated either until the final region $\lambda_n=\lambda_B$ is reached or because no successful trials were generated at some intermediate interfaces λ_i. Estimates for the conditional probability $P(\lambda_{i+1} \mid \lambda_i)$ at each interface are obtained, and an estimate of $P(\lambda_{n=B} \mid \lambda_0)$ is finally calculated as the product of these conditional probabilities:

$$P(\lambda_{n=B} \mid \lambda_0) = \prod_{i=0}^{n-1} P(\lambda_{i+1} \mid \lambda_i) = \frac{N_S^{(n-1)}}{\prod_{i=0}^{n-1} k_i} \tag{2}$$

which is the ratio of the total number of branches that eventually reach λ_n, to the total possible number of branches.[2,3] A new branching path is then generated by randomly choosing another point at λ_0 and following the same procedure outlined above to get a new estimate of $P(\lambda_{n=B} \mid \lambda_0)$. The final estimate of $P(\lambda_{n=B} \mid \lambda_0)$ is then obtained from the average over all such paths. For a complete description of the theoretical background of the algorithm, see Ref. 1.

B. FFS-LSE algorithm

The FFS-LSE protocol uses the *B*-committor probability (p_B) data obtained "on-the-fly" by the FFS-type simulation to obtain an estimate for the reaction

coordinate model. Accordingly, the p_B value for every interfacial point stored in the TPE trajectories, i.e., of $p_{B_j}^{\ i}$ for point *j* at λ_i, is obtained by using the following recursive equation:

$$p_{B_j}^{\ i} = \frac{1}{k_i} \sum_{m=1}^{N_j^i} p_{B_m}^{\ i+1} \qquad , i=n\text{-}1,n\text{-}2\ldots,0 \tag{3}$$

where N_j^i is the number of points reaching λ_{i+1} from point *j* at λ_i. Equation (3) is used starting from each point at λ_n where $p_{B_j}^{\ n} = 1$; then each point at $\lambda_{n\text{-}1}$ has $p_{B_j}^{\ n-1} = N_j^{n-1} / k_{n-1}$; each point at $\lambda_{n\text{-}2}$ has $p_{B_j}^{\ n-2} = \frac{1}{k_{n-2}} \sum_m p_{Bm}^{n-1}$; and so on back to *A*.

Once a FFS-type simulation is complete, $p_{B_j}^{\ i}$ values are obtained and all *m* candidate collective properties (suspected to be meaningful order parameters) are evaluated for every interfacial point stored. Because p_B is the ideal reaction coordinate,[15] a good order parameter (i.e., the best estimate for the reaction coordinate) model will be one that is able to "fit" these p_B data satisfactorily. To find such a model, one assumes that p_B follows a mathematical relation that depends on a number *m* of candidate collective variables (*q*):

$$\lambda(q) = p_B(q) = \sum_{k=1}^{m} \beta_k q_k + q^T \mathbf{A} q + \beta_0 + \varepsilon \tag{4}$$

where the parameters β_j, j=0,1,…,*m*, are the regression coefficients and absorb the units from the collective variables. The β_0 parameter allows the reaction coordinate to shift so the transition states are located at $\lambda(q)$=½. Interactions between collective variables are also included by use of the cross quadratic term in Eq. (4), where **A** is a matrix of adjustable parameters. The unknown coefficients in Eq. (4) are then found by standard least-square estimation (LSE) and the statistically significant terms in the

model are found by analysis of variance. The readers are referred to Ref. 3 for a detailed description of the FFS-LSE method.

C. Adaptive λ staging optimization algorithm

C.1 Interfaces 1 through n. The optimization algorithm seeks to allocate the computational effort of a FFS simulation to reduce the statistical error (per simulation period) with which the reaction rate constant $k_{A\to B}$ is estimated by optimizing for the position of interfaces $\{\lambda'\}$, which in turn results in the minimization of the variance in the $P(\lambda_{n=B} \mid \lambda_0)$ estimate with the constraint that:

$$P(\lambda_{n=B} \mid \lambda_0) = \prod_{i=0}^{n-1} P(\lambda_{i+1} \mid \lambda_i) \tag{5}$$

must remain constant. This optimization procedure leads to a net constant flux of partial trajectories between interfaces and hence a constant flux of connected paths throughout the region between the two end stable states per simulation period:

$$M_i P(\lambda_{i+1} \mid \lambda_i) = P(\lambda_{n=B} \mid \lambda_0)/(\alpha N_0) = N_s^{(i)} = N_s = \text{constant.} \tag{6}$$

where M_i is the total number of trial runs fired at interface λ_i and α is a Lagrange multiplier used to enforce Eq. (5). The readers are referred to Ref. 4 for a detailed description of the derivation of Eq. (6). This equation states that for optimal sampling, the $P(\lambda_{i+1} \mid \lambda_i)$ values [which are determined by the $\{\lambda\}$ set] must be set to attain a net constant flux of partial trajectories between interfaces $N_s^{(i)} = N_s$. Note that this equation does not fully specify the $P(\lambda_{i+1} \mid \lambda_i)$ values since we could simultaneously change the $P(\lambda_{i+1} \mid \lambda_i)$ and M_i values to satisfy it. This freedom allows us to externally input a desirable distribution of $P(\lambda_{i+1} \mid \lambda_i)$ values, e.g., a uniform distribution

with $P(\lambda_{i+1} \mid \lambda_i) = \left[P(\lambda_n \mid \lambda_0)\right]^{1/n}$. To do this, we use a function f of $P(\lambda_{i+1} \mid \lambda_i)$ that provides a one-to-one correspondence between an f value and a λ value [Eq. (40) in Ref. 4]. Such a function allows us to go from any prescribed $P(\lambda_{i+1} \mid \lambda_i)$ values to the sought-after λ values.

For $M_i P(\lambda_{i+1} \mid \lambda_i)$ to remain constant, k_i also has to be chosen such that $k_i = 1/P(\lambda_{i+1} \mid \lambda_i)$ for $0 < i < n$ while k_0 is chosen so that $N_s = k_0 P(\lambda_1 \mid \lambda_0)$ is fixed to the desired number of partial paths between interfaces. In summary, by tracking the conditional probabilities of reaching subsequent interfaces in λ space we can identify the "bottlenecks" of the FFS-type simulation and concentrate the sampling on these regions.

C.2 Interface 0. The location of the interface at λ_0 must be chosen such that the ensemble of stored configurations is uncorrelated and distributed over all the phase space sampled by the characteristic A→B pathways so that the ensemble of states at λ_0 is not under-sampled (if this happens, errors will propagate through the next interfaces). For this purpose, let y denote an observable property other than λ, whose values can be taken as providing a measure of phase space change that is distinct (e.g., "orthogonal") to that provided by λ. To estimate the correlation for a set of N measurements of the observable y for states at λ_0, we can define the autocorrelation function:

$$\mathrm{ACF(lag)} = \sum_{i=1}^{N-lag} \frac{\left(y_i - \bar{y}\right)\left(y_{i+\mathrm{lag}} - \bar{y}\right)}{\sum_{i=1}^{N}\left(y_i - \bar{y}\right)^2} \tag{7}$$

where $\bar{y}$ is the average for the complete run, N is the total number of stored point at λ_0, and the lag is the separation between stored states (in units of number of consecutive states at λ_0). This autocorrelation function can be intuitively understood as

an indicator of how *y* changes between consecutive stored states at λ_0. Note that from a single simulation in region A, the values of *y* are collected for *all* the stored states (the system must return to A between consecutive states) at different λ_0 values so that ACF can be determined for each λ_0. Of course, a suitable choice of the *y* property is crucial, and alternative definitions of the ACF [other than Eq. (7)] could be used that exploit some special symmetry in the behavior of the *y* property selected.

The ACF in Eq. (7) is expected to decay exponentially; i.e., $\mathrm{ACF(lag)} \propto \exp(-\mathrm{lag}/\tau_{\lambda_0})$, [see Fig. 3.9A in Sec. IV B], where τ_{λ_0} provides a measure of the autocorrelation time (in configurational space) at λ_0. More specifically, $m\,\tau_{\lambda_0}$ indicates the number of successive states that have to reach λ_0 before attaining (and storing) an essentially uncorrelated configuration. Here, *m* is a factor to tune the degree of uncorrelation desired; e.g., if *m*=2.3 then after $2.3\,\tau_{\lambda_0}$ crossings at λ_0 the ACF is decreased to 10% of its originally value. The average simulation time required between consecutive points reaching λ_0 is $\Delta t_{\lambda_0} = 1/\Phi_{A,0}$. Therefore, τ^*, the simulation time required to obtain an uncorrelated state at λ_0 is approximated by $\tau^* \propto \tau_{\lambda_0} \times \Delta t_{\lambda_0}$ and is the quantity we should aim to minimize since it would allow us to produce the maximum number of uncorrelated points at λ_0 for a given simulation time invested. From a single, long simulation run at basin A, the value of τ^* (which strongly depends on the position of λ_0) can be obtained for a broad range of pre-chosen discrete λ_0 values. The optimum location of λ_0 can then be readily determined by finding the minimum of the τ^* versus λ_0 curve (e.g., see Fig. 3.9B in Sec. IV B).

D. Stochastic thermostat for the FFS-MD simulations

For the FFS-MD scheme, the stochastic component required to achieve distinct trajectories was incorporated by controlling the temperature using an adaptation of the Andersen thermostat[22] (referred to as thermostat A). A similar approach has been used

before for generating stochastic trajectories when using MD with other path sampling methods.(8,14) For the starting conformation at $\lambda_{i,}$ as well as those for which thermostatting was required (approximately every four MD steps), velocities of all particles were assigned from a Gaussian distribution with a varying seed and standard deviation $\sigma_i = \sqrt{kT/m_i}$, where m_i is the mass of atom *i, T* is the system temperature and *k* is the Boltzmann constant. We evaluated the performance of thermostat A with respect to Nose-Hoover[22] temperature control, by performing exploratory MD runs on a 24.8 Å box containing 512 TIP3 water molecules, for which periodic boundary conditions were used. Fig. 2.3A shows a good agreement between the center of mass velocity distribution of water molecules for both thermostats (*t*=1 ns), indicating sampling of the appropriate (canonical) ensemble. In addition, the velocity autocorrelation function (VACF(t), truncated at 2 ps) for both thermostats is plotted in Fig. 2.3B, where an analogous decorrelation behavior is evident. An estimate of the self-diffusion coefficient (*D*) for each MD-thermostat run was also obtained from the VACF; the value of *D* found using thermostat A is 10% (absolute deviation) larger than the one found using Nose-Hoover thermostat. This deviation is roughly equivalent to the expected statistical variation of the estimate for *D* (which lacks long term correction), approximately 0.06 $Å^2$/ps. It is thus expected that the dynamical properties of the system will be reasonably close to those attained via conventional MD. To further elucidate the potential influence of the stochastic temperature control on the dynamic

properties of the system, supplementary TIP3 water MD runs as well as explicit solvent alanine dipeptide FFS-MD runs were performed using another version of our thermostat (thermostat B) with reduced stochastic perturbation as compared to thermostat A, equivalent to decreasing the coupling constant in the conventional Andersen thermostat. In particular, thermostat B lacks the initial perturbation enacted

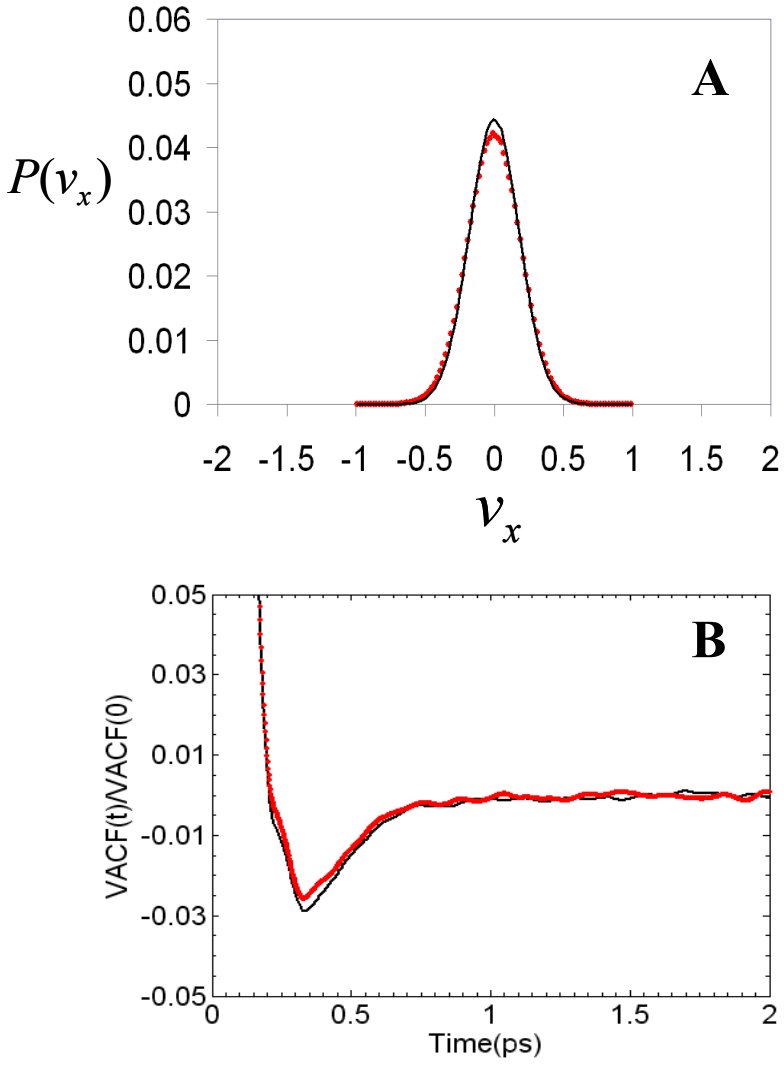


Figure 2.3. (*A*) Distribution for the center of mass velocity of water molecules for MD simulations using: (–) thermostat A and (•) Nosé-Hoover thermostat. (*B*) Time progression of the velocity autocorrelation functions (VACF) for thermostat A (–) and (•) Nosé-Hoover thermostat.

for thermostat A at each interface limit. For thermostat B, the estimate of *D* deviates only 2% from that using the Nose-Hoover thermostat. However, deficient sampling of phase space was apparent during preliminary explicit solvent alanine dipeptide FFS-MD simulations, where transition times were found to be 1-2 orders of magnitude higher than the expected values and a limited dispersion of paths was observed. As a result, thermostat A was considered to be the most appropriate for our goals and used throughout the FFS-MD simulations presented in this study. The reported exploratory analysis evidences the trade-off between the method's sampling efficiency and its accuracy. As a way for easing the latter limitation, ongoing efforts are focused on improving the thermostat by fine tuning the number and type of atoms being perturbed and the frequency of such perturbations.

III. SIMULATION DETAILS

All simulations were performed at 300 K in a parallel environment via CHARMM version c32b2,[18] using CHARMM all-atom force field. The leapfrog Verlet algorithm was used with a time step of 2 fs, and the SHAKE[22] algorithm was implemented for fixing the hydrogen bond length. The ψ dihedral angle was used as initial guess for the order parameter (λ) in both vacuum and explicit solvent FFS-type simulations. For maximum efficiency, parallelization of the FFS simulation was performed by computing each partial trajectory starting from a particular interface λ_i on a separate CPU (i.e., $N_s^i k_i$ simulations running at the same time).

A. Alanine Dipeptide in Vacuum

Fig. 2.4 shows the free energy landscape projected on the space of the ψ and ϕ torsion angles for the blocked (acetylated N-terminus, N-methylamide C-terminus) alanine dipeptide at 300K in vacuum, obtained from a 5 ns MD Replica Exchange

(REM)[22,23] simulation, spanning the 300-1000 K temperature range. The Weighed Histogram Analysis Method (WHAM)[24] was implemented for increased accuracy of the statistics. An analogous free energy landscape was obtained by using the MC algorithm in a run of length 10^6 MC cycles. The $C7_{eq}$ and C5 basins of interest are shown in Fig. 2.4, bounded by $50 < \psi < 100$ and $-100 < \phi < -65$, and by $150 < \psi < 195$ and $-135 < \phi < -165$, respectively, for negative values of ϕ. Basins of attraction A and B were selected so as to lie close to the minima obtained for states $C7_{eq}$ and C5 at 300 K, respectively. Accordingly, we defined the initial state as $\lambda_A \leq 80$ and the final state as $\lambda_B \geq 150$. The $\boldsymbol{\lambda}$ phase space between these stable states was partitioned using n=3 interfaces positioned at λ_i ($0 \leq i < n$): $\boldsymbol{\lambda}(x) = \{100,115,135\}$. The location of λ_0 was determined as described in Sec. II C using the ϕ angle as the variable measuring decorrelation between stored states at λ_0. The flux term in Eq. (1), $\Phi_{A,0}^{MC/MD}$, was obtained by averaging various MC/MD straightforward runs in the region A. Each flux term estimate is given by $\Phi_{A,0}^{MC/MD} = N_0 / \Gamma$, and obtained from Γ=5 ns straight forward MC/MD simulations, counting the number of times that the trajectory reached the first interface (λ_0) coming from A (i.e., N_0). The calculations were carried out using the BG method and the number of trials per point at λ_i was $k_i = 10$ ($0 \leq i < n$). Nine additional dihedral angles [including: θ, ϕ, and ζ (see Fig. 2.1)] were calculated for each of the interfacial stored points and used as possible collective variables for the reaction coordinate model estimation.

The FFS-MC and FFS-MD simulations were carried out as a series of blocks, each one consisting of $N_0 = 100$ points at λ_0. Each block consisted of a series of BG runs starting from a randomly selected configuration at λ_0 from which a branched path was generated and then used to estimate committor probabilities p_B. The optimal $\boldsymbol{\lambda}$ phase staging was determined from the first 200 BG simulations and is given in Table 2.2. This optimal staging was then employed for the following 500 simulations, which

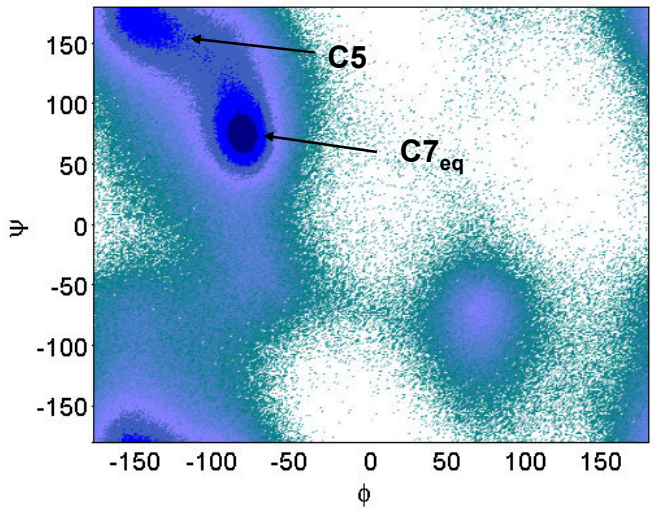


Figure 2.4. Free energy landscape for blocked alanine dipeptide in vacuum at 300 K. The color scheme for the visited states changes from highest (green) to lowest (gray/blue) elevations.

were used for the evaluation of $P(\lambda_{n=B}|\lambda_0)^{MC/MD}$. The p_B history data was obtained over all 700 blocks.

A.1 *MC Scheme.* The MC simulations were performed using the MC module in CHARMM, sampling states from a canonical distribution via the Metropolis acceptance criterion.[25] The optimized move set includes heavy atom translation, as well as rotation of hydrogen atoms, methyl group and main dihedral angles. The optimized move set, move frequency (weights) of atom groups and move sizes per MC step (i.e., cutoffs and parameter values) employed during FFS-MC sampling closely resemble those previously used by Hu et al.[25] for the MC study of alanine dipeptide in vacuum (see Table 2.1). In fact, our move set differs only in that linked moves (corresponding to move groups 7-10 in the study by Hu et al.[25] and equivalent to 0.7% of the total move weight) were left out of the scheme, given that their contribution to sampling efficiency was found to be negligible. Table 2.1 also lists the relevant parameters for the automatic optimization of move sizes [Acceptance Ratio Method (ARM) and Dynamically Optimized MC (DOMC)], which limit the changes for each move to yield a target Metropolis acceptance rate.[25] Dissimilar trajectories were achieved by changing the seed of the random number generator for the conformers at each interface. The value of the ψ angle (i.e., order parameter), needed to check if the trial path reached either the initial region or the next interface, was calculated every six MC moves (~2 ψ moves) rather than after each MC move to speed up the FFS-MC simulations.

A.2 *MD Scheme.* For the implementation of the FFS-MD scheme, the stochastic component was incorporated via an adaptation of the Andersen thermostat (thermostat A), as discussed in Sec II D. The velocities for all intermediate states (between interfaces) were adjusted as necessary to maintain the temperature close to the desired

value of 300 K. The ψ angle was calculated every two MD steps to check if the trial path reached either the initial region or the next interface.

B. Alanine Dipeptide in Explicit Solvent

An alanine dipeptide molecule was studied using the CHARMM all-atom force field with CMAP term correction.[26] The system was prepared by first solvating the peptide in TIP3 water within a cubic box with 24.8 Å on a side. Water molecules within a distance of 2.8 Å from the peptide were removed, leaving a total of 498 water molecules (compressed to 1g/ml). The system was then minimized with 1000 steps of a steepest descent algorithm. Periodic boundary conditions were applied using CHARMM's CRYSTAL facility, with a cutoff of 12 Å for non-bonded interactions. The conformation obtained after 10 ps of equilibration was used as initial state for several MD simulations performed towards the calculation of $\Phi_{A,0}$. Treatment of long range electrostatic interactions was undertaken via Particle Mesh Ewald.[22]

Fig. 2.5 shows the free energy landscape projected in $\psi - \phi$ dihedral angle space for alanine dipeptide in explicit solvent at 300 K. The contour was obtained by implementing the WHAM and REM under the same conditions indicated in Sec III A. The states of interest, β_2/α_R and $C5/C7_{eq}$, are bounded by $-55 < \psi < 30$ and $-45 < \phi < -195$, and by $120 < \psi < 195$ and $-35 < \phi < -205$, respectively. Both the faster $\beta_2/\alpha_R \Rightarrow C5/C7_{eq}$ and the slower $C5/C7_{eq} \Rightarrow \beta_2/\alpha_R$ transitions were explored by comparing the transition times obtained with other available estimates and formulating appropriate reaction coordinate models for each reaction. For the faster transition, regions A and B were chosen as having a ψ angle value close to the minima for states β_2/α_R and $C5/C7_{eq}$ at 300 K (see Fig. 2.5). Accordingly, the initial and final states were defined by taking $\lambda_A \leq 20$ and $\lambda_B \geq 130$, respectively, for negative values of ϕ. For the reverse transition (slower), regions A and B were defined by taking $\lambda_A \geq 130$ and $\lambda_B \leq 20$,

Table 2.1. Optimized move set for MC simulation in vacuum.[25] Parameters for the automatic optimization of move sizes (ARM and DOMC) are also given.

Move Description	Instances	WEIGHT(%)	ARM P_t	DOMC F
Heavy atom anisotropic translation	10	15.0	0.20	2.0
Hydrogen atom rotation	6	6.6	0.20	4.5
Methyl group rotation	3	33.3	0.25	9.0
ϕ rotation	1	7.3	0.45	4.5
ψ rotation	1	32.8	0.50	0.5
θ rotation	1	2.5	0.55	2.5
ζ rotation	1	2.5	0.55	2.5

Table 2.2. Optimized {λ} sets for vacuum and explicit solvent FFS-MC and FFS-MD simulations.

i	**Vacuum** $C7_{eq} \Rightarrow C5$ transition			**Explicit solvent**			
	Initial {λ} set	**MC**	**MD**	**$\beta_2/\alpha_R \Rightarrow C5/C7_{eq}$ [faster transition]**		**$C5/C7_{eq} \Rightarrow \beta_2/\alpha_R$ [slower transition]**	
		Optimized λ staging	Optimized λ staging	Initial {λ} set	Optimized λ staging	Initial {λ} set	Optimized λ staging
0	100	100	100	30	30	120	120
1	115	108	108	60	50	105	104
2	135	117	117	90	74	90	88
3	$\lambda_{n=B}$ = 150	150	150	105	108	60	58
4				$\lambda_{n=B}$ =130	130	$\lambda_{n=B}$ =20	20

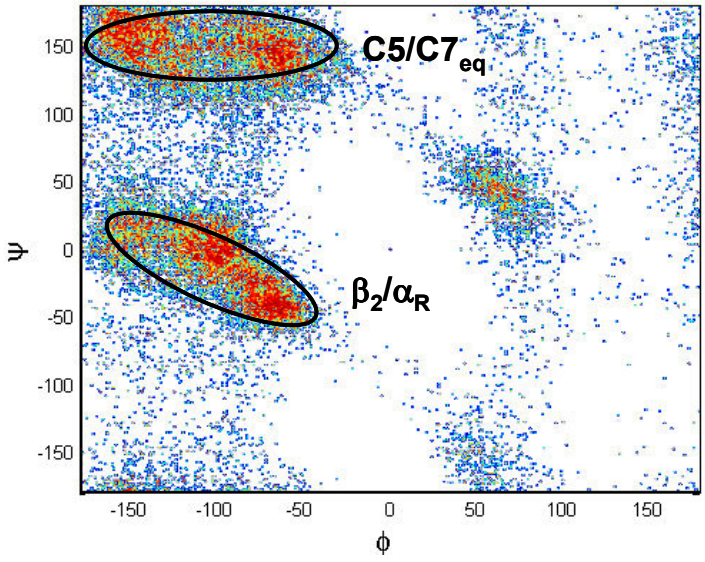


Figure 2.5. Free energy landscape for alanine dipeptide in explicit solvent at 300 K.

respectively, for negative values of ϕ (see Fig. 2.5). The $\boldsymbol{\lambda}$ space was partitioned using n=4 interfaces positioned at λ_i ($0 \le i < n$): $\boldsymbol{\lambda}(x) = \{30,60,90,105,130\}$ for the faster transition, and $\boldsymbol{\lambda}(x) = \{120,105,90,60,20\}$ for the slower transition. The location of λ_0 was determined by measuring the autocorrelation function (ACF) [i.e., Eq. (7)] for all the states at λ_0 along the ϕ angle (the results are discussed in Sec. IV B). The flux term in Eq. (1), $\Phi_{A,0}$, was obtained by averaging various MD runs in region A as explained in Sec. III A. The calculations were carried out using the BG method and the number of trials per point at λ_i was $k_i = 5$ ($0 \le i < n$), with starting points randomly sampled from inside the region A. In addition to ψ, the other three main dihedrals θ, ϕ, and ζ were calculated for each stored conformation reaching consecutive interfaces. As possible collective variables for the reaction coordinate model estimation, we also calculated the distance between atoms 2H and $2C_\beta$ as well as the electrostatic torque around bond 1C-2N from solvent forces on atom 3H. As determined by Ma and Dinner,[16] these two collective variables appear to capture the solvent's role in the isomerization reaction (see Fig. 5(a) in Ref.15). Following the mentioned study, the force vector between the water molecules and the atom 3H, $\vec{F}_{H_2O,3H}^{elec} = -\nabla_{3H}\sum_i E_{i,3H}^{elec}$, is used to calculate the electrostatic torque around bond 1C-2N from solvent forces on atom 3H: $\Gamma_{1C-2N}^{3H} = (\vec{F}_{H_2O,3H}^{elec} \times \vec{r}_{2N-3H}) \bullet \hat{e}_{1C-2N}$. $E_{i,3H}^{elec}$ is the Coulomb electrostatic energy between atom i and solute atom 3H and the sum runs over all solvent atoms. For further information regarding these two order parameters see Ma and Dinner.[16]

The FFS-MD simulations were carried out as a series of blocks, as described in Sec. III A. The optimal $\boldsymbol{\lambda}$ phase staging for the faster and slower transitions was determined from the first 100 BG simulations for each reaction and is given in Table 2.2. This new staging was then employed for the following 200 simulations of each

transition, which were used for the evaluation of $P(\lambda_{n=B}|\lambda_0)$. The p_B history data was obtained over all 300 blocks.

IV. RESULTS

A. Alanine dipeptide in vacuum

The equivalence between the MC and MD approaches was assessed by comparing the values obtained for the transition time during the evaluated simulation period. For this purpose, a time step was estimated for a MC move of the ψ dihedral angle (Δt_{MC}) by using an average diffusion coefficient ($\overline{D}_{MD}$) calculated from the average standard deviation ($\overline{\sigma}_{MD}$) of the ψ dihedral angle and time step (Δt_{MD}) of the MD runs performed for the calculation of $\Phi_{A,0}^{MD}$. The expression for the MC time step is then:

$$\Delta t_{MC} = 0.33 * \frac{2\sigma_{MC}^{\ 2}}{\overline{D}_{MD}} \quad \text{, where } \overline{D}_{MD} = \frac{2\overline{\sigma}_{MD}^{\ 2}}{\Delta t_{MD}} \tag{8}$$

The 0.33 prefactor in the Δt_{MC} expression corresponds to the overall probability of occurrence of a ψ angle move within the complete MC move set (see WEIGHT column on Table 2.1).

Using Eq. (1) with values for the average fluxes $\langle\Phi_{A,0}^{MC}\rangle = 1.43*10^{12}$ s^{-1} and $\langle\Phi_{A,0}^{MD}\rangle = 1.12*10^{12}$ s^{-1}, and $P(\lambda_{n=B} \mid \lambda_0)^{MC} = 0.161 \pm 0.03$ and $P(\lambda_{n=B} \mid \lambda_0)^{MD} = 0.223 \pm 0.025$, the transition times found in our MC and MD simulations for the $C7_{eq} \Rightarrow C5$ reaction were 4.5 ± 0.84 ps and 4.05 ± 0.46 ps, respectively. The proximity of these results supports the advocated correspondence between both FFS-type methods. Furthermore, the kinetic transition time ($1/k_{A\to B}$) obtained for this reaction is consistent with the ones estimated by Chun et al.[27] and Vedell and Wu[28] between these two basins. The former study calculated a 2.7 ps

transition time both via atomistic MD simulations and a rigid body MBO(N)D[27] method using CHARMM, while the latter estimated a transition rate between 0.1-3 ps using a multiple shooting algorithm with a MOIL based force field. In this system (and for solvent explicit case of IV.B), we expect that our implementation of thermostat A introduces an additional error in our rate constant estimation and so the standard deviations given above are likely underestimated; however, such an error is expected to be much smaller than that which would be introduced by sub-sampling the stochastic trajectories with a weakly-coupled thermostat (see Sec. II.D).

The ψ and ϕ dihedral angles are commonly considered to give a good characterization of the conformational diversity of this peptide in vacuum (i.e., serve as reaction coordinates). To investigate the extent to which another collective variable might be important, we applied the FFS-LSE method to obtain a good estimate for the reaction coordinate model. It is important to stress that in addition to the four main torsion angles shown in Fig. 2.1, we considered six other angles as physical variables which could provide relevant information. Such coordinates were included so as to also assess the effect of interaction terms between these variables. The p_B history was obtained from the TPE by the method outlined in Sec. II B and fitted to a tentative regression model, including the ten torsion angles and quadratic interaction terms between them. The analysis of variance (ANOVA) for this model indicated that the terms for ψ and ϕ are the only significant ones. The insignificance of the additional dihedral angle terms in the model evidences their limited contribution to the characterization of transition pathways and of the studied stable states [see Fig. 2.6B and 2.7B]. Following this outcome, a second LSE was performed considering only the ψ and ϕ regressors in the reaction coordinate model (see Tables 2.3 and 2.4), from which we obtained:

$$\lambda(\psi,\phi) = p_B = \left[-2.25 + 0.71(\psi) - 1.09(\phi) + 0.16(\psi^2) - 0.05(\phi^2) + 0.33(\psi\phi)\right], \quad (9)$$

$$\lambda(\psi,\phi) = p_B = \left[-0.83 + 0.31(\psi) + 0.18(\psi^2) + 0.11(\phi^2) + 0.13(\psi\phi)\right] \quad (10)$$

for the FFS-LSE-MC [Eq. (9)] and MD [Eq. (10)] simulations. The dihedrals (ψ and ϕ) are in radians. Tables 2.3 and 2.4 report a small *P*-value for the *F* statistic of the model, indicating that the reaction coordinate models of Eqs. (9) and (10) describe the variability of the p_B data with statistical significance. As expected from the fact that it correlates strongly with the committor probability distribution, both simulations consistently indicate that ψ by itself is capable of predicting the p_B isocommittor surface (i.e., larger *F*-value for terms involving the ψ angle regressor). The quadratic terms (i.e., ψ^2 and ϕ^2) in the reaction coordinate model are necessary to capture the curvature followed by the connecting pathways between stable states on the p_B isocommittor surface. Even though the coefficients for the reaction coordinate models obtained from the two independent BG simulations are different, Figures 2.6B and 2.7B show similar λ response surfaces for the optimal order parameters from Eqs. (9) and (10) projected onto the ψ and ϕ free-energy landscape. In both cases, the reaction coordinate model identifies the TS dividing surface [$\lambda(\psi,\phi) = p_B(\psi, \phi) = ½$], passing through $\psi \approx 125°$. This value matches the ψ value observed at the top of the energy barrier in Figs. 2.6A and 2.7A for FFS-MC and FFS-MD simulations, respectively. Moreover, the optimized order parameter expressed as the p_B iso-committor surface captures the A and B boundary regions. Because the model's surface is not bound to lie within the p_B interval [0,1], all states with a $p_B \leq 0$ can be enclosed together, defining the initial basin of attraction *A*. Likewise, the region *B* is defined by enclosing all states with $p_B \geq 1$. Additional FFS-LSE iterations could be implemented where in each cycle the current optimized order parameter is used to obtain a new estimate for the reaction coordinate until convergence; in this case, just one iteration was enough to get suitable results.[5]

Table 2.3. LSE parameters and analysis of variance for the reaction coordinate model of the FFS-MC simulation in vacuum. The ψ and ϕ angles are given in radians.

Source	Sum of Squares	df	Coefficient [β_i]	Mean Square	F-value	P-value
Model	6072.1	5		1214.4	68057	< 0.0001
ψ	209.3	1	0.71	209.3	11727	< 0.0001
ϕ	92.1	1	-1.09	92.1	5160	< 0.0001
ψ^2	2314.6	1	0.16	2314.6	129711	< 0.0001
ϕ^2	3.47	1	-0.05	3.47	194.4	< 0.0001
$\psi\,\phi$	180.1	1	0.33	180.1	10095	< 0.0001
Constant			-2.25			
Residual	1639.0	91850		0.018		
Corr. Total	7711.0	91855				
R^2	0.78					

Table 2.4. LSE parameters and analysis of variance for the reaction coordinate model of the FFS-MD simulation in vacuum. The ψ and ϕ angles are given in radians.

Source	Sum of Squares	df	Coefficient [β_i]	Mean Square	F-value	P-value
Model	6039.74	4		1659.9	75341	< 0.0001
ψ	63.0	1	0.31	63.0	2857	< 0.0001
ψ^2	3086.2	1	0.18	3086.2	140077	< 0.0001
ϕ^2	56.1	1	0.11	56.1	2544	< 0.0001
$\psi\,\phi$	43.8	1	0.13	43.8	1990	< 0.0001
Constant			-0.83			
Residual	2090.6	94888		0.022		
Corr. Total	8730.3	94892				
R^2	0.76					

Overall, we can state that the FFS-LSE approach is capable of selecting a combination of physical collective variables involving degrees of freedom (like ϕ) that by themselves are not clearly relevant to the reaction. Our results indicate that the ψ and ϕ order parameters were found to be sufficient for predicting the dynamic pathways of the transition. Moreover, an interaction term between these variables and quadratic terms is found to be necessary for a more complete description of the isocommittor surface curvature.

B. Alanine dipeptide in water

The slower C5/C7$_{eq}$ $\Rightarrow$ β_2/α_R reaction in the presence of water molecules has been analyzed in numerous studies, some of which have estimated the transition time and/or determined an appropriate order parameter that describes this transition.[15,16] For our FFS-MD simulations, the transition time obtained was 328 ± 62 ps, with $<\Phi_{A,0}>$ = 2.70×10^{11} s^{-1} and $P(\lambda_{n=B} \mid \lambda_0) = 0.0118 \pm 0.002$. Other studies have reported results of the same order of magnitude. For example, Bolhuis et al.[15] estimated a rate constant of 10 ns^{-1} (equivalent to a 100 ps transition time) using TPS with AMBER 94 force field, whereas Chekmarev et al.[10] calculated a mean first passage time of 249 ps using Brownian dynamics and an Analytic Generalized Born with Nonpolar Interactions (AGBNP) implicit solvent model with OPLS-AA force field.

With respect to the appropriate order parameter for the slower isomerization, the spatial distribution of configurations belonging to the TS ensemble over the free-energy landscape projected onto the ψ and ϕ angles indicates that relative to its behavior in vacuum, the transition dynamics in solution is highly diffusive; this suggests that additional variables capturing the role of solvent dynamics in the transition are necessary for the reaction coordinate model.[15,16] This is also evident in

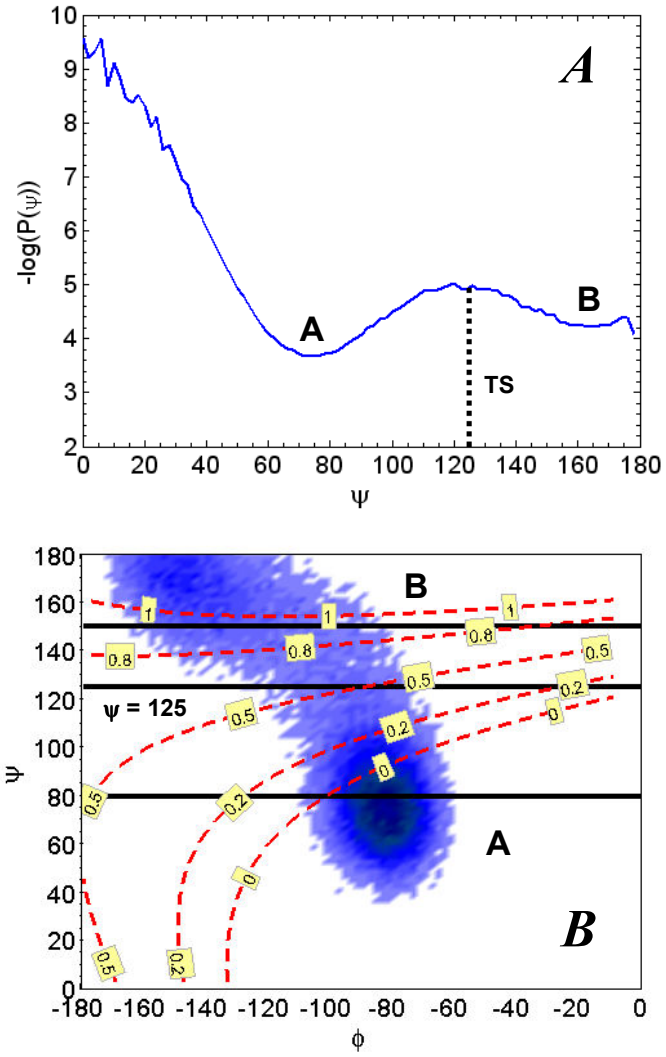


Figure 2.6. Results for the FFS-MC simulations in vacuum at 300 K. (*A*) Free energy profile along the ψ dihedral angle as order parameter. The dotted line corresponds to the value of $\psi \approx 125°$ at the transition state. (*B*) Contour of the free energy landscape ($\psi - \phi$ plane). The color scheme for the visited states changes from highest (gray/light blue) to lowest (black/dark blue) elevations. The solid (black) lines correspond to the initial order parameter $\lambda = \psi$: 80 (state A upper limit), 125 (TS), and 150 (state B lower limit). The dotted (red) lines correspond to the $\lambda = p_B$ isocommittor surface.

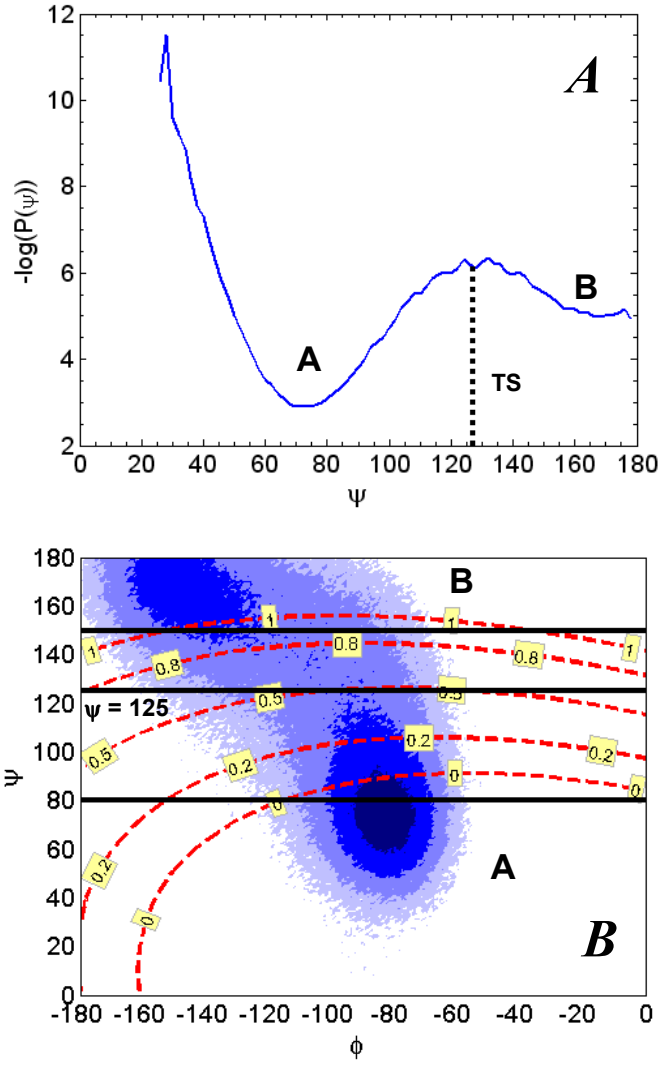


Figure 2.7. Results for the FFS-MD simulations in vacuum at 300 K. (*A*) Free energy profile along the ψ dihedral angle as order parameter. (*B*) Contour of the free energy landscape over the ψ-ϕ plane. Color and line schemes are the same as those indicated in the caption of Fig. 2.6.

the broadness of the committor probability distribution reported by Bolhuis et al.,[15] for configurations at the top of the free energy barrier along the ψ angle as order parameter. The ψ and ϕ dihedrals are not sufficient for an adequate representation of the isomerization dynamics because a good reaction coordinate should lead to a distribution of p_B for the TS ensemble peaked at ½.[29] Ma and Dinner[16] determined that the coupling of additional solvent collective variables to the principal reaction coordinate variable (i.e., ψ and/or ϕ) is necessary to obtain an appropriate dynamic behavior of this transition. These authors analyzed 1132 physical variables that describe solute-solvent and solvent-solvent interactions, finding that the best reaction coordinate model is composed of three descriptors: the ψ angle, the distance between atoms 2H and 2C$_\beta$ $\left|r_{\mathbf{2H\text{-}C_\beta}}\right|$, and the electrostatic torque around bond 1C-2N from solvent forces on atom 3H ($\Gamma^{3H}_{1C\text{-}2N}$). In this work, we used the FFS-LSE method to test the significance of these descriptors and of the interaction terms between them.

The p_B history was obtained from the TPE by the method outlined in Sec. II B and fitted to a tentative regression model, including collective variables for the four main dihedral angles, $\left|r_{\mathbf{2H\text{-}C_\beta}}\right|$, $\Gamma^{3H}_{1C\text{-}2N}$, and quadratic interaction terms between these variables, to obtain:

$$\lambda(\psi,\phi) = p_B = [1.30 - 1.38(\psi) - 0.005(\phi) + 0.03(\left|r_{2H-C_\beta}\right|) + 0.35(\psi^2) + 0.0002(\Gamma^{3H}_{1C-2N})^2] \quad (11)$$

where the ψ and ϕ angles are given in radians. $\left|r_{\mathbf{2H\text{-}C_\beta}}\right|$ and $\Gamma^{3H}_{1C\text{-}2N}$ are given in Å and kcal/mol units, respectively. Table 2.5 shows the LSE parameters and ANOVA for the reaction coordinate model of this reaction. The *P*-value in Table 2.5 for the *F* statistics of the model [Eq. (11)] is very small, indicating that at least one of the six variables has a nonzero regression coefficient. The upper portion of Table 2.5 also gives the

LSE of each parameter, the partial *F*-value statistic, and the corresponding *P*-value. As expected, the partial *F*-test shows that the ψ angle is the most important collective variable (i.e., ψ and ψ^2) for the description of the model. Furthermore, our results confirm that the variables that describe the solvent dynamics during the transition (i.e., $\left|r_{\mathbf{2H-C_\beta}}\right|$ and $\Gamma^{\mathbf{3H}}_{\mathbf{1C-2N}}$ regressors, which have P-value < 0.05 for the partial *F*-test statistics) are necessary for a good estimate of the order parameter.

Fig. 2.8 shows a map of the probability density (P_{TPE}) of finding a configuration (ψ, ϕ) in the transition path ensemble (TPE) after a long FFS run [i.e., $P_{\mathrm{TPE}}(\psi, \phi)$ is incremented by one if a trajectory connecting A (C5/$C7_{eq}$) and B (β_2/α_R) visits this configuration at least once], where it can be seen that the phase space sampling of the trajectories connecting the two stable states A and B is comparable to that observed for this region during the REM simulations (see Fig. 2.5).

We placed λ_0=120 < λ_A=130 so that the ensemble of states at λ_0 is not underestimated. The location of λ_0 was determined by running a simulation in the region A and then calculating the ACF [i.e., Eq. (7)] for the states at λ_0 along ϕ angle. Fig 2.9A shows the ACF(lag) for states collected at λ_0=110, 115, 120, 125, and 130 as a function of the separation between stored states (i.e., lag). Fig 2.9B shows τ_{λ_0} and the average simulation time required between stored points at λ_0 (i.e., Δt_{λ_0}) as a function of λ_0. Hence, practically uncorrelated states at λ_0 (e.g., ACF=0.1) are obtained after $2.3\,\tau_{\lambda_0} = 7$ configurations have crossed λ_0=120 between stored states and the simulation time required to obtain an uncorrelated state is proportional to $\tau_{\lambda_0} \times \Delta t_{\lambda_0} = 6.76$ ps. This choice of λ_0 maximizes the number of uncorrelated starting points (at λ_0) for a given simulation time, and those points produce a uniform distribution of states along ϕ as shown in Fig. 2.8 by the red region along the horizontal line at ψ=120.

Table 2.5. LSE parameters and analysis of variance for the reaction coordinate model of the slower C5/C7$_{eq}$ $\Rightarrow$ β_2/α_R reaction of an FFS-MD simulation in explicit solvent. The ψ and ϕ angles are given in radians. $\left|r_{2H-C_\beta}\right|$ and $\Gamma^{3H}{}_{1C\text{-}2N}$ are given in Å and kcal/mol, respectively.

Source	Sum of Squares	df	Coefficient [β_i]	Mean Square	F-value	P-value
Model	481.7	5		96.34	4510	< 0.0001
ψ	94.87	1	-1.38	94.87	4441	< 0.0001
ϕ	0.1778	1	-0.005	0.1778	7.9	0.0050
$\left\|r_{2H-C_\beta}\right\|$	0.2493	1	0.03	0.2493	11.7	0.0006
ψ^2	26.67	1	0.35	0.1434	1248	< 0.0001
$(\Gamma^{3H}{}_{1C\text{-}2N})^2$	0.1434	1	0.0002		6.7	0.0097
Constant			1.30			
Residual	60.74	2843		0.021		
Corr. Total	542.43	2848				
R^2	0.90					

Regarding the faster $\beta_2/\alpha_R \Rightarrow C5/C7_{eq}$ isomerization in explicit solvent, the average flux was found to be $<\Phi_{A,0}>=8.40\times10^{11}$ s^{-1} and $P(\lambda_{n=B} \mid \lambda_0) = 0.04 \pm 0.008$, resulting in a transition time of 31 ± 6 ps. The value found is consistent with previous studies that have analyzed this reaction; be.g., Oliveira et al.[19] estimated a mean escape time of 80 ps using accelerated MD simulations with AMBER force field in explicit solvent, whereas Chekmarev et al.[10] used the AGBNP implicit solvent model with the OPLS-AA force field to obtain a mean first passage time of 27 ps. Moreover, West et al.[17] studied the reaction via Milestoning (a path sampling scheme) using MOIL package, finding a mean first passage time of 64 ps.

A reaction coordinate analysis was also performed for the faster transition. Table 2.6 shows the LSE parameters and ANOVA for the reaction coordinate model. The *P*-value for the partial *F* statistic indicates that the ψ and ϕ dihedral angles and the variables that describe the solvent dynamics during the transition (i.e., $\left|r_{2H-C_\beta}\right|$ and $\Gamma_{1C-2N3H}$) are necessary for a good estimate of the order parameter. The partial *F*-test also shows that the ψ angle is the most important collective variable for the description of the model. A linear regression involving the four descriptors provides a reasonably complete description of the p_B data. Ma and Dinner[16] also found reasonable accuracy for their p_B data using a linear regression, but involving only three descriptors. Conversely, the FFS-LSE procedure with the p_B database yielded a reaction coordinate model which includes an extra term for the ϕ dihedral angle. The predicted transition state isocommittor surface (λ=p_B=0.5) on the ψ and ϕ plane depends on the Γ^{3H}_{1C-2N} variable, as seen on Fig. 2.10. The latter result, underlining the important role played by the solvent dynamics in the transition, is consistent with the behavior observed by other researchers[15,16] where the committor probability distribution of the ensemble of configurations existing at the top of the free energy barrier (along $\psi = 60^o$) was found to be quite broad.

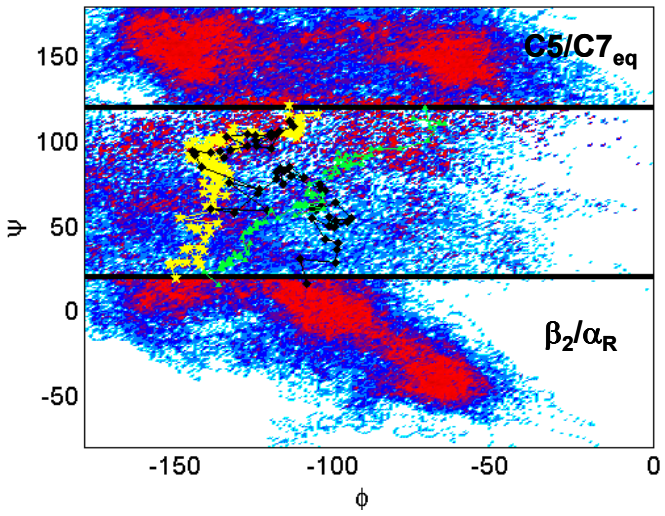


Figure 2.8. Density map (P_{TSE}) obtained from the TPE for several FFS-MD runs for the $C5/C7_{eq} \Rightarrow \beta_2/\alpha_R$ reaction at 300K in explicit solvent. The color scheme for the visited states changes from most (red) to least (light blue) visited region. The solid (black) lines correspond to: λ_0=120 and $\lambda_{n=B}$=20. Three representative trajectories are also shown.

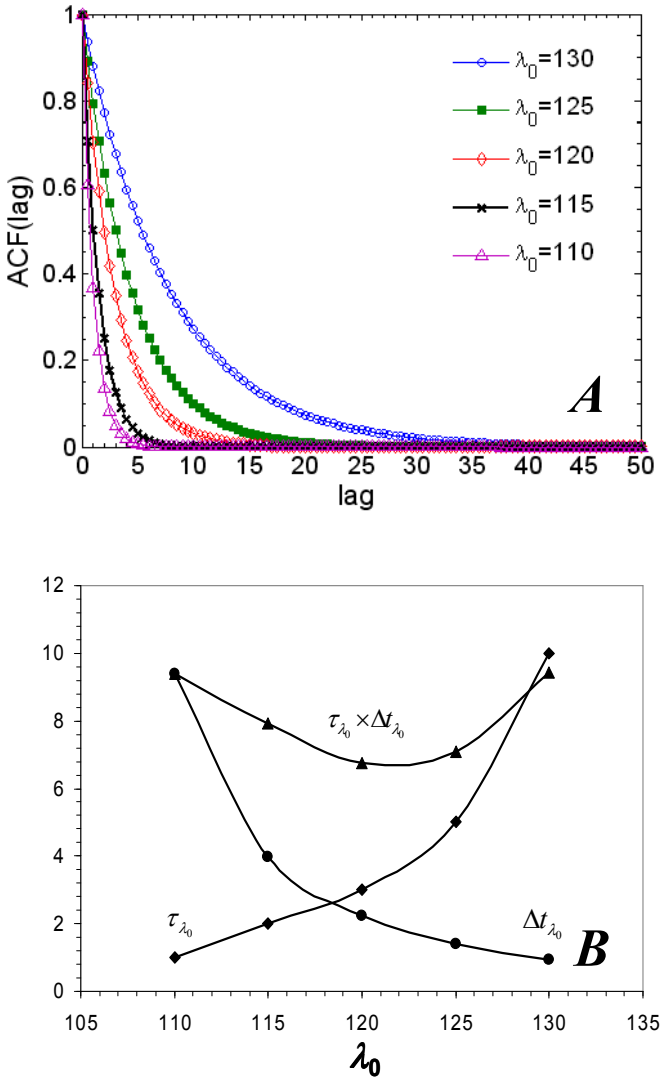


Figure 2.9. Results for the optimization process of the λ_0 positioning in the FFS-MD simulation for the C5/C7$_{eq}$ $\Rightarrow$ β_2/α_R reaction at 300K in explicit solvent: (*A*) Auto correlation functions for the ϕ angle for states collected at λ_0= 130, 125, 120, 115 and 110, and (*B*) (♦) τ_{λ_0}, (•) Δt_{λ_0} (picoseconds), and (▲) $\tau^* = \tau_{\lambda_0} \times \Delta t_{\lambda_0}$ (picoseconds) curves as a function of the location of λ_0.

Table 2.6. LSE parameters and analysis of variance for the reaction coordinate model of the faster $\beta_2/\alpha_R \Rightarrow C5/C7_{eq}$ reaction from a FFS-MD simulation in explicit solvent. The ψ and ϕ angles are given in radians. $\left|r_{2H-C_\beta}\right|$ and $\Gamma^{3H}{}_{1C\text{-}2N}$ are given in Å and kcal/mol, respectively.

Source	Sum of Squares	df	Coefficient [β_i]	Mean Square	F-value	P-value
Model	338.4	4		84.61	4083	< 0.0001
ψ	317.7	1	0.68	317.7	15331	< 0.0001
ϕ	0.2656	1	0.003	0.2656	12.8	0.0003
$\left\|r_{2H-C_\beta}\right\|$	1.98	1	-0.07	1.98	95.7	0.0001
$\Gamma^{3H}{}_{1C\text{-}2N}$	0.0902	1	-0.003	0.0902	4.35	0.0371
Constant			-0.37			
Residual	104.9	5064		0.021		
Corr. Total	443.4	5068				
R^2	0.80					

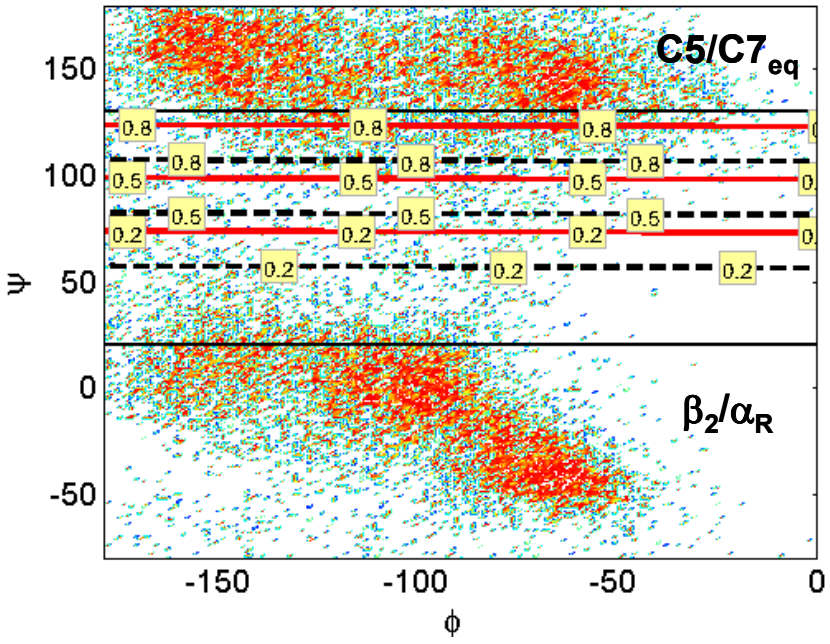


Figure 2.10. Isocommittor surfaces obtained during the FFS-MD simulations for the $\beta_2/\alpha_R \Rightarrow C5/C7_{eq}$ reaction at 300K in explicit solvent. The solid (black) lines correspond to: λ_A=20 and $\lambda_{n=B}$=130. The color scheme changes from highest (light blue) to lowest (red) elevations. The isocommittor surfaces $\lambda=p_B$ (see Table 2.6) are shown for fixed values of $\left|r_{2H-C_\beta}\right| = 3$ Å and $\Gamma^{3H}{}_{1C\text{-}2N}$=30 kcal/mol (solid red lines), and for fixed values of $\left|r_{2H-C_\beta}\right| = 3$ Å and $\Gamma^{3H}{}_{1C\text{-}2N}$=-30 kcal/mol (dotted black lines). For the p_B data considered, $\left|r_{2H-C_\beta}\right| = 3$ Å is the average value observed, and $\Gamma^{3H}{}_{1C\text{-}2N}$ = [-30, 30] are the [lower, upper] limits of the range of values observed.

V. CONCLUSIONS

The main aim of the current paper was to demonstrate the use of FFS and address some its shortcomings, by simulating a well-known testbed system in continuum via MD. To this end, we studied the $C7_{eq} \Rightarrow C5$ transition of alanine dipeptide in vacuum using FFS-MC and FFS- MD simulations, as well as the forward and reverse $\beta_2/\alpha_R \Leftrightarrow C5/C7_{eq}$ transitions for the same peptide in explicit solvent. Transition rate constant values were determined from both FFS- MC and FFS-MD simulations for the vacuum reaction, and from FFS-MD simulations for the explicit solvent reactions. The good agreement between the rate constant values obtained from both FFS-type simulations in vacuum, as well as the consistency between our results and those found in the literature[10,15,17,19] for both vacuum and explicit solvent simulations validate the use of the FFS-MD approach for the study of biomolecular transitions. Moreover, successful FFS applications such as the present one, which use a widely tested force field/simulation package in parallel, open the door for more challenging, larger scale applications.

For the systems studied in this work, we also showed that the FFS-LSE[5] algorithm, in combination with our new proposed method to optimize the position of the λ_0 interface and an adaptive algorithm to optimize the position of subsequent λ interfaces, gave a good estimate of the order parameter. Analogous isocommittor surfaces (i.e., reaction coordinate models) were obtained from the FFS-MC and FFS-MD simulations in vacuum. Moreover, the reaction coordinate model for the $\beta_2/\alpha_R \Rightarrow C5/C7_{eq}$ transition in explicit solvent obtained from the FFS-LSE method confirms previous results from Ma and Dinner,[16] further validating our FFS-MD approach. Our results show that the FFS-LSE method is successful in identifying an optimal order parameter and that using a single variable (i.e., ψ angle) as order parameter is not sufficient to describe the committor probability distribution. From the analysis of the

p_B model, other significant interaction terms were identified between the physical variables describing the solute and solvent dynamics. Overall, our results provide additional mechanistic insights on the dynamics of alanine dipeptide (i.e., interaction terms between the common variables used to describe the transitions) both in vacuum and in explicit solvent. Having used a consistent force field and simulation protocol, our results also unambiguously illustrate the differences in the transition kinetics of peptides with and without solvent. The advocated approach provides a promising platform for future studies of biomolecular transitions.

ACKNOWLEDGMENTS

The authors are grateful for support from the National Science Foundation Award 0553719 and from an ACS-PRF grant.

REFERENCES

1. R. J. Allen, D. Frenkel, and P. R. ten Wolde, J. Chem. Phys. **124**, 194111 (2006).

2. R. J. Allen, D. Frenkel, and P. R. ten Wolde, J. Chem. Phys. **124**, 024102 (2006).

3. R. J. Allen, P. B. Warren, and P. R. ten Wolde, Phys. Rev. Lett. **94**, 018104 (2005).

4. E. E. Borrero and F. A. Escobedo, J. Chem. Phys. **125**, 164904 (2006).

5. E. E. Borrero and F. A. Escobedo, J. Chem. Phys. **127**, 164101 (2007).

6. E. E. Borrero and F. A. Escobedo, J. Chem. Phys. **129**, 024115 (2008).

7. C. Dellago and P. G. Bolhuis, in *Atomistic Approaches in Modern Biology: from Quantum Chemistry to Molecular Simulations* (Springer-Verlag Berlin, Berlin, 2007), Vol. 268, pp. 291.

8. C. Dellago and P. G. Bolhuis, in *Advanced Computer Simulation Approaches for Soft Matter Sciences III* (Eds. C. Holmer, K. Kremer, Springer, 2008), Vol. 221, pp. 167.

9. J. Juraszek and P. G. Bolhuis, Biophys. J. **95**, 4246 (2008).

10. D. S. Chekmarev, T. Ishida, and R. M. Levy, J. Phys. Chem. B **108**, 19487 (2004).

11. C. L. Brooks and D. A. Case, Chem. Rev. **93**, 2487 (1993).

12. T. Lazaridis, D. J. Tobias, C. L. Brooks, and M. E. Paulaitis, J. Chem. Phys. **95**, 7612 (1991).

13. B. M. Pettitt and M. Karplus, Chem. Phys. Lett. **121**, 194 (1985).

14. D. J. Tobias and C. L. Brooks, J. Phys. Chem. **96**, 3864 (1992).

15. P. G. Bolhuis, C. Dellago, and D. Chandler, Proc. Nat. Acad. Sci. U. S. A. **97**, 5877 (2000).

16. A. Ma and A. R. Dinner, J. Phys. Chem. B **109**, 6769 (2005).

17. A. M. A. West, R. Elber, and D. Shalloway, J. Chem.Phys. **126,** 145104 (2007).

18. B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus, J. Comput. Chem. **4**, 187 (1983).

19. C. A. F. de Oliveira, D. Hamelberg, and J. A. McCammon, J. Chem. Phys. **127**, 175105 (2007).

20. D. Moroni, T. S. van Erp, and P. G. Bolhuis, Phys. A - Stat. Mech. & Appl. **340**, 395 (2004).

21. T. S. van Erp, D. Moroni, and P. G. Bolhuis, J. Chem. Phys. **118**, 7762 (2003).

22. D. Frenkel and B. Smit, *Understanding Molecular Simulation: From Algorithms to Applications*, 2nd ed. (Academic, Boston, 2002).

23. D. J. Earl and M. W. Deem, Phys. Chem. Chem. Phys. **7**, 3910 (2005).

24. A. M. Ferrenberg and R. H. Swendsen, Phys. Rev. Lett. **63**, 1195 (1989).

25. J. Hu, A. Ma, and A. R. Dinner, J. Comput. Chem. **27**, 203 (2006).

26. A. D. Mackerell, J. Comput. Chem. **25**, 1584 (2004).

27. H. M. Chun, C. E. Padilla, D. N. Chin, M. Watanabe, V. I. Karlov, H. E. Alper, K. Soosaar, K. B. Blair, O. M. Becker, L. S. D. Caves, R. Nagle, D. N. Haney, and B. L. Farmer, J. Comput. Chem. **21**, 159 (2000).

28. P. Vedell and Z. Wu, Int. J. Num. Anal. & Mod. (submitted) (2008).

29. C. Dellago, P. G. Bolhuis, and P. L. Geissler, Adv. Chem. Phys. **123**, 1 (2002).

CHAPTER 3

KINETICS AND MECHANISM OF THE UNFOLDING N-L TRANSITION OF TRP-CAGE IN EXPLICIT SOLVENT VIA OPTIMIZED FORWARD FLUX SAMPLING SIMULATIONS*

I. INTRODUCTION

Trp-cage is a model protein that has been extensively used in computational research due to its small size, the presence of secondary and tertiary structures in its folded state and its fast two-state folding. A number of studies have been reported in the literature that use Molecular Dynamics (MD),[1-2] enhanced MD techniques[3-8] and other methods[9-13] to identify the energy landscape, the folding rate, and the main conformations pertaining the folding-unfolding pathway. Of particular interest are those studies that have focused on the kinetics and folding mechanism of Trp-cage employing rare event sampling techniques such as Transition Path Sampling ,[14] Transition Interface Sampling (TIS),[15] and Forward Flux Sampling (FFS).[16-17] In a recent publication, Juraszek and Bolhuis[3,18] calculated rate constant values for the native to loop (hereupon N-L) transition and proposed a reaction coordinate model for this system. These authors proposed that the N-L transition is the rate limiting step of the main route (N-L-U) from the native to the unfolded (U) state, and that an alternate route consisting of two intermediates (N-P_d-I-U) had low occurrence. While the L-N rate constant obtained via TIS is close to the experimental k_{UN} value[19] (56% higher), the values reported for the reverse N-L transition using TIS and FFS are respectively about an order of magnitude higher and lower than those found in experiment for the N-U transition.[19] Moreover, there was a considerable difference between the conformational spaces sampled from both TIS and FFS methods. Juraszek and Bolhuis

* C. Velez-Vega, E.E. Borrero, and F. A. Escobedo, J. Chem. Phys. (in press)

pointed out that the Transition Path Ensemble (TPE) harvested by FFS concentrated in a non-representative region of the phase space, resulting in the overestimation of the free energy barrier and hence underestimation of the rate constant.[18] In general, compared to other interface-based path sampling methods, FFS is likely more sensitive to the choice of the order parameter used to partition the phase space and has more difficulty relaxing the pathways in directions orthogonal to the imposed order parameter.[20-21]

Recently, we highlighted the importance of optimizing the implementation of FFS for the study of rare events in complex systems,[22-23] making use of various strategies that tune the position of the interfaces such that an efficient sampling of the path space is obtained. The optimum staging for a given order parameter is then used to generate committor probability data via the FFS-LSE[22] method to obtain an estimate for the reaction coordinate of the system. This procedure can be iterated to improve the reaction coordinate model. Because the Trp-cage transitions have been extensively studied by several path sampling methods, it is an ideal candidate to test the performance of optimized FFS methods. Consequently, the main goal of this paper is twofold: (i) to validate the usefulness of optimized FFS techniques for studying the kinetics and mechanism of biomolecular transitions, by quantifying the N-L unfolding kinetics of the Trp-cage mini-protein and (ii) to further elucidate the mechanism of this transition by analyzing the TPE obtained from our FFS simulations.

II. METHODS

In this section, a brief overview is given of the FFS methodology for sampling the Transition Path Ensemble (TPE) and the calculation of the average transition rate constants $k_{A\to B}$ (Sec. II A). In Secs. II B and C, we summarize the methods that can optimize the spacing of λ, and the novel protocol for the FFS simulations adopted in

this work. In Secs. II D and E we briefly describe our method for selection of the order parameter (λ) as reaction coordinate, and the programmed trajectory termination algorithm used for reducing the computational cost of our simulations. Finally, we describe the stochastic thermostat[18,24] (Sec. II F) adopted to allow the implementation of FFS with deterministic MD simulations.

A. Forward Flux Sampling (FFS) via conventional Branched Growth (BG).

In BG the phase space is partitioned by employing a series of (n+1) nonintersecting interfaces defined by an order parameter λ and whose values increase monotonically as the interfaces come close to region B. Starting from a randomly selected configuration at the first interface λ_0, a branched "tree" is generated by harvesting partial paths that connect successive interfaces λ_i ($0 \le i \le n$-1). At each interface (i), multiple trial runs (k_i) are performed per point to promote $N_S^{(i)}$ successful partial paths between interfaces [$\lambda_i \rightarrow \lambda_{i+1}$].

The average rate of transitions $k_{A\rightarrow B}$ from two well-defined states A and B is estimated by using an "effective positive flux" expression:[16-17,25-26]

$$k_{A\rightarrow B} = \overline{\Phi}_{A,0} P(\lambda_{n=B} \mid \lambda_0) \tag{12}$$

The $\overline{\Phi}_{A,0} = N_0 / \Gamma$ term (i.e., the total average flux of trajectories from A to λ_0) is estimated by counting the number N_0 of effective positive crossing events of the first interface at λ_0 from a simulation of length Γ in the initial basin A. The $P(\lambda_{n=B} \mid \lambda_0)$ probability that a trajectory reaching λ_0 from A will reach to B without returning to A is estimated from the product of conditional probabilities $P(\lambda_{i+1} \mid \lambda_i)$ to jump from one interface to the next:

$$P(\lambda_{n=B} \mid \lambda_0) = \prod_{i=0}^{n-1} P(\lambda_{i+1} \mid \lambda_i) = \prod_{i=0}^{n-1} \frac{N_S^{(i)}}{M_i} \tag{13}$$

where M_i is the total number of trial runs fired at each λ_i. For a complete description of the algorithm, see Ref. 16.

B. Adaptive λ staging optimization algorithm

The efficiency of the FFS simulations depends, among other things, on the number and position of the interfaces and on how extensively different interfaces are sampled.[23,25] Adaptive algorithms have been proposed to optimize the λ sampling for both the number and position of the interfaces (i.e., optimized λ phase staging).[23]

Optimizing the ensemble at interface λ_0

If the ensemble of states at λ_0 is under-sampled by generating just a few uncorrelated starting points at λ_0, the TPE harvested for the FFS simulation may represent only a small portion of the phase space relevant to the transition, leading to erroneous transition rate constants. Hence, suitable positioning of the first interface λ_0 is crucial for FFS efficiency. This under-sampling behavior is more evident when a poor order parameter is used to partition the λ phase. To overcome this issue, we recently proposed a protocol to store an ensemble of uncorrelated configurations distributed over the whole phase space sampled by the characteristic pathways.[24] For this purpose, we evaluated the time $\tau^* \propto \tau_{\lambda_0} \times \Delta t_{\lambda_0}$ required to obtain an uncorrelated state at λ_0. $\Delta t_{\lambda_0} = 1/\overline{\Phi}_{A,0}$ is the average simulation time required to reach consecutive points at λ_0 [$\overline{\Phi}_{A,0}$ was defined in Eq. (12)]. The constant τ_{λ_0} provides a measure of the autocorrelation time at λ_0 and is determined by measuring the autocorrelation function (ACF) at λ_0, i.e., $\mathrm{ACF(lag)} \alpha \exp(-\mathrm{lag}/\tau_{\lambda_0})$, where *lag* is the separation between stored states (in units of number of consecutive states at λ_0).

The ACF is calculated by the correlation of a set of N measurements of y for states at λ_0:

$$\text{ACF(lag)} = \sum_{i=1}^{N-lag} \frac{\left(y_i - \bar{y}\right)\left(y_{i+\text{lag}} - \bar{y}\right)}{\sum_{i=1}^{N}\left(y_i - \bar{y}\right)^2} \tag{14}$$

where y is an observable property whose values provide a measure of phase space change that is ideally "orthogonal" to that provided by λ. In Eq. (14), $\bar{y}$ is the average for the complete run in basin A. One could find the optimum location of λ_0 by the minimum of the τ^* versus λ_0 curve (this curve can be constructed from data of a *single* run in basin A).

Optimizing the position of subsequent interfaces

An adaptive optimization algorithm has also been proposed which seeks to reposition the interfaces $\{\lambda\}$ to allocate the computational effort in a FFS simulation in such a way as to concentrate the sampling in the bottleneck regions.[23] In order to decrease the statistical error in the estimate of the rate constant, this optimization algorithm prescribes a net constant "forward" flux of partial trajectories between interfaces $N_s^{(i)} = N_s$:[23]

$$M_i P(\lambda_{i+1} \mid \lambda_i) = P(\lambda_{n=B} \mid \lambda_0)/\alpha = N_s^{(i)} = N_s = \text{constant.} \tag{15}$$

Now, for $M_i P(\lambda_{i+1} \mid \lambda_i)$ to remain constant, Ref. 23 suggests to freely specify $P(\lambda_{i+1} \mid \lambda_i)$ and let the total number of trial runs M_i fired at λ_i to be chosen such that the condition in Eq. (15) is achieved. In this work, we targeted a uniform distribution for $P(\lambda_{i+1}|\lambda_i)$; i.e.,

$$P(\lambda_{i+1} \mid \lambda_i) = \left[P(\lambda_n \mid \lambda_0)\right]^{1/n} = \text{constant}, \qquad i=1,2,..,n\text{-}1 \tag{16}$$

The new optimized $\{\lambda'\}$ staging that corresponds to these values of $P(\lambda_{i+1} \mid \lambda_i)$ is found from a special "interpolating" function $f(\lambda)$ [Eq. (40) in Ref. 23] constructed from the existing $P(\lambda_{i+1} \mid \lambda_i)$ vs. λ data.

The reader is referred to Ref. 23 for a detailed description of the adaptive optimization algorithm method.

C. Constrained Branched Growth (CBG) Scheme

With the conventional BG method, it is very difficult to enact the constant-flux condition of Eq. (15) as even small fluctuations in the expected $P(\lambda_{i+1} \mid \lambda_i)$ produce a fast growth of partial paths as one approaches state B.[16,25] In the new CBG method, a constant number of trial runs M_i are fired at each interface *i*, just as in the original or "direct" FFS (DFFS) method.[16] In principle, those M_i trials could be randomly sampled from the $N_s^{(i-1)}$ points that reached λ_i, in which case the CBG would essentially be a special case of the DFFS where only one starting point is used at λ_0. However, this scheme may not be the most effective for cases when $N_s^{(i-1)}$ is large and a minimum number of trial runs per state is desired. We therefore used in this work a slight variant of this idea where L_i states are randomly selected from the set of $N_s^{(i-1)}$ trajectories that successfully reached λ_i ($L_i \leq N_s^{(i-1)}$), and k_i^j points ($1 \leq j \leq L_i$) are started from each of these states, now aimed at λ_{i+1}. The value of L_i is selected such that:

$$L_i = \begin{cases} \dfrac{M_i}{k_{\min}}, & for \quad N_s^{(i-1)} > \dfrac{M_i}{k_{\min}} \\ N_s^{(i-1)}, & for \quad N_s^{(i-1)} \leq \dfrac{M_i}{k_{\min}} \end{cases} \tag{17}$$

where k_{min} corresponds to the minimum number of shots per point (here set to $k_{min} \geq 4$). Consequently, k_i^j is given by:

$$k_i^j \approx \frac{M_i}{L_i} \qquad for \quad j \leq L_i \tag{18}$$

This procedure has a threefold purpose: (i) to fix M_i such that the condition of a constant flux of partial trajectories between interfaces (N_s) of Eq. (15) can be more readily achieved, (ii) to avoid the uncontrolled growth of the computational cost of the BG simulation by making the number of trial runs independent of the number of states reaching λ_i, and (iii) to set a minimum value of k_i^j ($k_i^j \geq k_{min}$) such that committor probabilities with two significant digits can be obtained for the FFS-LSE method (see Sec. II.D).

An alternate BG procedure denoted "random" BG or RBG was used for our initial set of FFS runs with the "unoptimized" reaction coordinate; RGB is also capable of preventing a disproportionate growth of partial paths and is described in Appendix C. The efficiency and accuracy of RBG and CBG with respect to the conventional BG method were tested with a simple model. These results (detailed in Appendix C) show that CBG is the most robust of these methods and was therefore implemented for our second "optimized" set of FFS simulations leading to the rate constant calculation and reaction coordinate analyses presented in Sec. IV (Results and Discussion).

D. FFS-LSE algorithm

The B-committor probability (p_B) is presumably the ideal reaction coordinate of the system.[27] Recently, we proposed a method denoted FFS-LSE, that extracts p_B data "on-the-fly" from BG simulations. For each interfacial point stored in the TPE trajectories, an estimate for the p_B value is obtained by recursively calculating:[23]

$$p_{B_j}^{\ i} = \frac{1}{k_i^j} \sum_{m=1}^{N_j^i} p_{B_m}^{\ i+1}, \qquad i=n\text{-}1,n\text{-}2,\ldots,1 \tag{19}$$

where $p_{B_m}^{\ i+1}$ is the committor probability to reach B for each point m at λ_{i+1} that connects with state j at λ_i and k_i^j is the number of trial runs fired for the point j at λ_i.

The p_B history data collected over the whole phase space region connecting states A and B is then fitted to a mathematical relation that depends on any number of candidate collective variables (suspected to be meaningful order parameters for the system's dynamics). Standard least-square estimation (LSE) and an analysis of variance (ANOVA) are used to find the statistically significant terms in the model. This new estimate model could then be used to partition the phase space, in a second iteration of the entire process, to generate additional p_B data to refine (i.e., LSE-fit) the order-parameter model. The reader is referred to Ref. 22 for a detailed description of the FFS-LSE method.

E. Programmed Trajectory Termination

A procedure inspired by the pruning algorithm[16] was implemented for further increasing the efficiency of our simulations, in which trajectories from λ_i which have not yet reached λ_{i+1} or λ_A after time τ_T are terminated, and their potential contribution to $P(\lambda_{i+1}|\lambda_i)$ accounted for by a transition probability, P_T. Appropriate values of τ_T can be obtained from preliminary runs at interface $\lambda_{N\text{-}1}$, in which the time required for trajectories to reach either the initial or the final state is measured; values at the higher end of the resulting time distribution are good choices for τ_T (in this study we used 5 ps). The conditional probability at each interface i is then modified as:

$$P(\lambda_{i+1}|\lambda_i) = P_c(\lambda_{i+1}|\lambda_i) + P_T N_T / M_i \tag{20}$$

where $P_c(\lambda_{i+1}|\lambda_i)$ is the conventional probability of reaching λ_{i+1} from λ_i (in which only successful trajectories are included), N_T is the number of trajectories that are terminated when trying to reach λ_{i+1} or λ_A, and M_i is the total number of partial trajectories fired at λ_i. No trajectories are terminated before at least one trajectory reaches λ_B for a particular FFS simulation. Although the number of terminated trajectories throughout our runs was low and virtually limited to the last interfaces, termination of these trajectories significantly reduced the overall simulation time. The preliminary runs used for optimizing the λ staging were also used to estimate P_T; we set $P_T = 0.25$ based on the number of trajectories that effectively reached basin B from λ_{n-1} after τ_T.

This procedure should be especially useful for large systems with pathways displaying slow diffusion that may seriously hamper FFS efficiency.

F. Stochastic Thermostat for the FFS-MD simulations

Uncorrelated trajectories were achieved in the FFS-MD algorithm by performing temperature control via an Andersen thermostat (coupling constant τ = 0.02 ps), coupled with velocity reassignment at each λ interface as well as after every 200 MD steps. For this purpose, velocities were randomly generated from a Maxwell-Boltzmann distribution at the temperature of interest (300 K). Using a system of TIP3P water molecules, we have previously[24] studied the performance of an analogous version of our thermostat with respect to those conventionally employed to achieve deterministic behavior (e.g., using Nose-Hoover thermostat). In that work, we found that a 10% (absolute) deviation in the water self-diffusion coefficient (D) with respect to the one obtained using Nose-Hoover thermostat, was acceptable without noticeably affecting the behavior of the center of mass velocity distribution and the velocity autocorrelation function when compared to canonical simulations. Using the deviation

in *D* from Ref. 24 as our reference value, we simulated SPC water diffusion for three independent 1 ns MD simulations at 300 K using Nose-Hoover (τ = 0.02 ps), velocity rescaling (τ = 0.02 ps) and our thermostat. *D* was calculated for the three cases from the slopes of the Mean Square Displacement plots, after discarding the initial and final 100 ps of each run. The average values of *D* for the runs that employed velocity rescaling and our thermostat deviate respectively 4% ($4.01 \pm 0.08 \times 10^{-5}$ cm^2/s) and 8% ($3.84 \pm 0.17 \times 10^{-5}$ cm^2/s) from the average value obtained using Nose-Hoover thermostat ($4.18 \pm 0.20 \times 10^{-5}$ cm^2/s). Given that the deviation in *D* for this version of our thermostat is lower than the tolerance (10%), it is conjectured that its use will lead to a dynamic behavior consistent with that of conventional thermostats.

III. SYSTEM SETUP

Following the work by Juraszek and Bolhuis,[18] our MD runs and FFS-MD trajectories were evolved using the GROMACS simulation package[28] with the OPLS-AA force field and the SPC water model. The representative NMR structure of the folded Trp-cage miniprotein (PDB code 1L2Y) was solvated in a 4.58 nm box with 3077 water molecules. The system was then neutralized and energy-minimized using the Steepest Descent algorithm. To eliminate artifacts due to the initial water distribution, a 10 ps MD run was performed at 282 K with the position of the protein restrained. The system was then heated to 300 K via a 80 ps simulated annealing run. This was followed by a 100 ns equilibration run at 300 K and 1 bar using Andersen Thermostat and Berendsen pressure coupling. A step size of 2 fs was used for all our simulations.

A. Order Parameters

With the goal of directly comparing our results with those obtained by Juraszek and Bolhuis,[18] we considered the same order parameters that they proposed for system characterization. Throughout the simulations we monitored the α-carbon radius of gyration (r_{gyr}), fraction of native contacts (ρ), α-carbon RMSD from the native structure ($RMSD_{ca}$), RMSD of residues 2-8 from an ideal helix ($RMSD_{hx}$), RMSD of the hydrophobic core formed by tryptophan 6 and prolines 12 and 17-19 ($RMSD_{hc}$), protein's solvent accessible surface area ($SASA$), the number of water molecules within a 0.4 nm radius from the tryptophan (n_{wat}), and the distances between donors and acceptors in the H-bonds of the Asp-9 – Arg-16 salt bridge (sb_1, sb_2 and sb_3).

B. FFS Setup

Initial Staging

As mentioned in the previous section, we used $RMSD_{hx}$ as initial order parameter for the phase space partition between the two stable states A and B. With the intent of directly comparing our results with those reported by Juraszek and Bolhuis (18), we used $RMSD_{hx} = 0.05$ nm as the choice of λ_{A} for our initial set of FFS runs. Similarly for these runs, λ_0 and $\lambda_{\mathrm{B}} = \lambda_{\mathrm{n}}$ were located at 0.06 nm and 0.22, respectively. Ten interfaces were placed between λ_0 and λ_{B}, initially positioned at $\{0.07, 0.08, 0.09, 0.10, 0.11, 0.12, 0.14, 0.16, 0.18, 0.20\}$, $\lambda_0 < \lambda_{\mathrm{i}} < \lambda_{\mathrm{n}}$. Their location was tuned via the adaptive staging optimization algorithm (discussed shortly) to increase the efficiency of the TPE sampling.

Despite maintaining the location of $\lambda_0 = 0.06$ nm for our initial set of FFS runs, we employed the algorithm described in Sec. IIB to determine τ^* (i.e., the simulation

time required to obtain an uncorrelated state at λ_0) so that an ensemble of uncorrelated configurations at λ_0 could be collected from a simulation in basin A. Specifically, we determined:

$$lag_{\lambda 0} = m\,\tau_{\lambda_0} \tag{21}$$

where m = -ln[%ACF/100] is a factor to tune the degree of uncorrelation desired and $lag_{\lambda 0}$ indicates the number of successive states that have to reach λ_0 before attaining (and storing) an essentially uncorrelated configuration. For example, if m = 2.3 then after 2.3 τ_{λ_0} crossings at λ_0 the ACF is decreased to 10% of its originally value.

The autocorrelation function (ACF) [i.e., Eq. (14)] was calculated along $RMSD_{ca}$ and n_{wat} as possible variables measuring decorrelation between stored states at λ_0. Figure 3.1 shows the ACF(lag) for states collected at λ_0 = 0.06 nm as a function of the separation between stored states (i.e., lag); as expected, τ_{λ_0} depends on the y property [in Eq. (14)]. Thus, seemingly uncorrelated states at λ_0 (e.g., ACF = 0.1) are obtained after 2.3 τ_{λ_0} = 1150 and 460 configurations have crossed λ_0 = 0.06 nm between stored states when y is the $RMSD_{ca}$ and $n_{\text{wat,}}$ respectively. The longer (more conservative) estimate of $lag_{\lambda 0}$ given by $RMSD_{ca}$ is the safer choice. The average simulation time required between consecutive points reaching λ_0 is $\Delta t_{\lambda_0} = 1/\Phi_{A,0}$ = 38 ps (see Sec. IV).

Hence, we waited for τ_{λ_0} =1200 consecutive states crossing λ_0 =0.06 nm before storing an uncorrelated configuration, expecting to obtain an ensemble of states at λ_0 that is not underestimated. The simulation time required to obtain an uncorrelated state was ~ 2.3 $\tau_{\lambda_0} \times \Delta t_{\lambda_0}$ = 46 ns. Following the ideas described in Sec. IIB, we found that the optimum location of λ_0 (i.e., the minimum of the τ^* versus λ_0 curve) is between 0.075 and 0.085, but we kept λ_0 = 0.06 nm for consistency with the study in Ref. 18.

A pre-equilibrated structure was used as starting conformation for a 100 ns MD run at 300 K from which the effective positive flux $\Phi_{A,0} = 2.67\times10^4$ μs^{-1} and the initial structures at λ_0 were obtained.

As part of our FFS setup we optimized our $\{\lambda_i\}$ set via the adaptive λ staging optimization algorithm presented in Sec. IIB. Table 3.1 shows our optimized set $\{\lambda_i'\}$ for which $P(\lambda_{i+1} \mid \lambda_i) = P_\lambda \approx 0.33$ for $0 \le \lambda_i < 11$, as well as our initial staging and the one used by Juraszek and Bolhuis[18] in their FFS simulations. Note that the algorithm concentrates the location of the interfaces in the region close to basin A (i.e., in the uphill region leading to the transition state) such that all the partial trajectories starting at any particular λ_i have approximately the same probability of reaching the next interface at λ_{i+1} (listed in Table 3.1).

Staging for second set of FFS runs

The p_B history data obtained from our first set of FFS runs was used to find a first estimate for the reaction coordinate model (λ^{opt1}). The λ^{opt1} expression was then used as guiding order parameter for a second set of FFS runs aimed at improving our reaction coordinate estimate. For this purpose, the position of the λ_i^{opt1} interfaces was also optimized using the methodology presented in Sec. IIB. We defined basins A and B by taking $\lambda_A^{opt1} \le$ -0.6 and $\lambda_B^{opt1} \ge 1.0$, respectively; in Sec. IVB we discuss the practical considerations of this selection. The location of λ_0^{opt1} was determined by measuring the autocorrelation function (ACF) [i.e., Eq. (14)] for all the states at λ_0^{opt1} along y = $RMSD_{hx}$, obtained from our flux MD simulation in basin A. Figure 3.2 shows τ_{λ_0} and the average simulation time required between stored points at λ_0 (i.e., Δt_{λ_0}) as a function of λ_0. The choice of λ_0^{opt1} = -0.5 (i.e., minimum at the $\tau_{\lambda_0} \times \Delta t_{\lambda_0}$ curve) maximizes the number of uncorrelated starting points (at λ_0^{opt1}) for a given simulation time. Hence, we placed $\lambda_0^{opt1} = -0.5 < \lambda_A = -0.6$ so that the ensemble of states at λ_0^{opt1} is not underestimated, and uncorrelated states at λ_0^{opt1} are obtained

after $\tau_{\lambda_0} = 10$ configurations have crossed λ_0^{opt1} between stored states. Accordingly, a new effective positive flux $\Phi_{A,0} = 1.49\times10^4$ μs^{-1} was measured through λ_0^{opt1} such that the simulation time required to obtain an uncorrelated state is $\sim \Delta t_{\lambda_0} \times \tau_{\lambda_0} = 670$ ps. We then conducted a preliminary CBG run to optimize the $\{\lambda_i^{opt1}\}$ set via the adaptive optimization algorithm discussed in Sec. IIB. The initial $\{\lambda_i^{opt1}\}$ and optimized staging are listed in Table 3.2. The optimum staging was then used to obtain an improved estimate of the reaction coordinate model (λ^{opt2}).

IV. RESULTS AND DISCUSSION

Two sets of FFS runs were performed in this study. The first set consisting of 500 RBG runs used $RMSD_{hx}$ as λ order parameter, whereas the second set of 200 CBG runs was guided by the optimized reaction coordinate (λ^{opt1}) found from the first set of simulations via FFS-LSE.

For each FFS run, the starting conformation at λ_0 was randomly selected from an ensemble of structures collected as described in Sec. IIIB. Concerning the initial set of FFS runs, Figure 3.3 shows the phase space distribution of structures (red) at or close to λ_0 (0.06 nm) along order parameters $RMSD_{ca}$, $RMSD_{hc}$, r_{gyr} and n_{wat}, from which the ensemble was constructed. When compared to the sampling of the native basin at 300 K attained via Replica Exchange[29] (that used all pairs exchanges[30] and spanned the 250-450 K range), it is evident that the $\lambda_0 = 0.06$ nm states are well distributed over the equilibrium range of values of the order parameters considered[**]. Likewise for our second set of FFS runs, an appropriate distribution of the starting states at $\lambda_0^{opt1} = -0.5$ was verified. Accordingly, the trajectories generated via FFS

[**] It is noted that simulations performed via Replica Exchange starting at the native state (basin A, with $RMSD_{hx}$ and $RMSD_{ca}$ boundaries shown in Fig. 3.3A), and via MD starting at the loop state (basin B, with boundaries $0.23 < RMSD_{hx} < 0.43$ and $0.35 < RMSD_{ca} < 0.88$) show somewhat different envelopes for basins A and B when compared to those defined by Juraszek and Bolhuis in Table 3 of Ref. 18.

likely started from structures that are representative of the equilibrium phase space at λ_0 (initial FFS runs) or λ_0^{opt1} (second FFS runs).

The rate constant calculation presented below corresponds to that obtained using the second set of "optimized" FFS runs. Details on the analogous analysis performed for the initial set of FFS runs is discussed in Appendix D.

A. Estimation of the Rate Constant

From our second set of FFS simulations we obtained a rate constant value of k_{NL} = (7.6 μs)$^{-1}$ for the N-L transition. This value agrees well with the experimental unfolding rate of k_{NU} = (12.7 μs)$^{-1}$ obtained by Qiu *et al.*.[19] As pointed out in the introduction, direct comparison of these rate values is justified based on the observation of Juraszek and Bolhuis[3,18] that the N-L transition is the rate limiting step of the overall unfolding process. We note the existence of a highly probable "lower energy" pathway (for details see Sec. IVB) which contributes the majority of the successful trajectories for the rate constant calculation. In contrast, our rate constant value differs from the one found in Ref. 18 using an FFS scheme (k_{NL} = (100 μs)$^{-1}$). In that paper it was suggested that the discrepancy with respect to the experimental rate is partly due to the higher sensitivity of FFS, as compared to TPS or TIS, to the choice of λ. While FFS may indeed be more sensitive that way, the rate constant value obtained from our initial FFS runs ((8 μs)$^{-1}$, see Appendix D) indicates that using $RMSD_{hx}$ as order parameter λ for this transition is an effective way for sampling the TPE and attaining a suitable rate via FFS, provided that one uses an uncorrelated, well sampled λ_0 ensemble and an optimized staging. In Appendix D the results from our initial FFS simulations are directly compared to the FFS analysis carried out by Juraszek and Bolhuis[18] (a direct comparison is appropriate since both use λ = $RMSD_{hx}$).

Table 3.1. Initial, optimized and reference $\{\lambda\}$ sets for order parameter $\lambda = RMSD_{hx}$. The $P(\lambda_i|\lambda_{i+1})$ values found using our optimized set are also reported.

i	Initial $\{\lambda\}$ set	Optimized set		$\{\lambda\}$ set used in Ref. 18
		$\{\lambda_i'\}$	$P(\lambda_i\|\lambda_{i+1})$	
0	0.06	0.060	0.28	0.06
1	0.07	0.069	0.31	0.08
2	0.08	0.078	0.31	0.10
3	0.09	0.085	0.30	0.11
4	0.10	0.090	0.28	0.12
5	0.11	0.098	0.33	0.14
6	0.12	0.102	0.32	0.16
7	0.14	0.110	0.31	0.17
8	0.16	0.128	0.33	0.18
9	0.18	0.152	0.34	0.19
10	0.20	0.185	0.33	0.20
11(=n)	0.22	0.22	----	0.22

Table 3.2. Initial and optimized $\{\lambda'\}$ sets for the reaction coordinate model λ^{opt1}. The $P(\lambda_i|\lambda_{i+1})$ values found using our optimized set are also reported.

i	Initial $\{\lambda\}$ set	Optimized set	
		$\{\lambda_i'\}$	$P(\lambda_i\|\lambda_{i+1})$
0	-0.50	-0.50	0.28
1	-0.40	-0.41	0.29
2	-0.25	-0.14	0.27
3	-0.10	-0.03	0.28
4	0.00	0.12	0.27
5	0.20	0.28	0.27
6	0.40	0.43	0.28
7	0.60	0.62	0.28
8	0.80	0.82	0.25
9(=n)	1.00	1.00	----

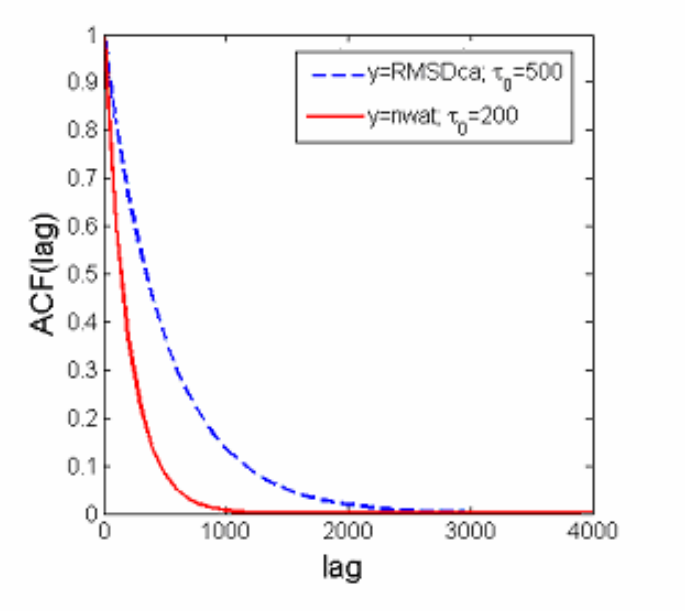


Figure 3.1. ACF(lag) for states at λ_0=0.06 nm as a function of the separation between stored states (lag), for y = $RMSD_{ca}$ and y = n_{wat}.

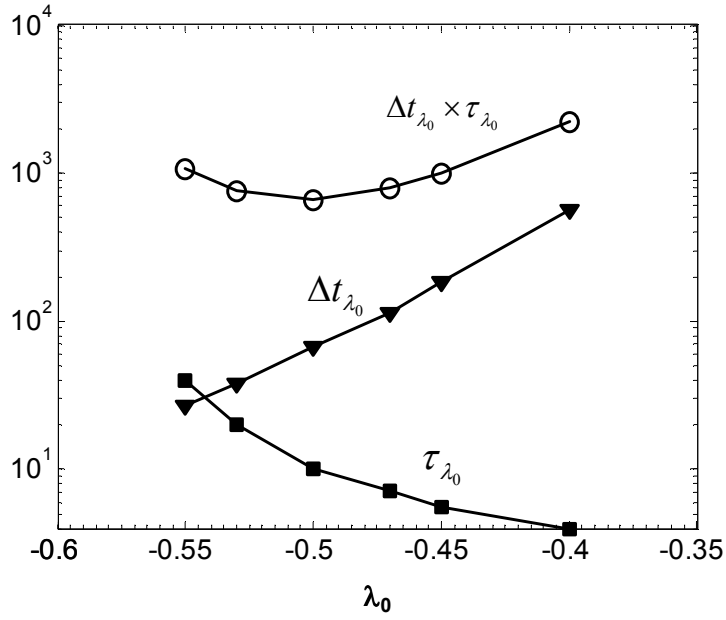


Figure 3.2. Optimization of the λ_0^{opt1} positioning [$RMSD_{hx}$ was used as the y parameter in Eq. (14)]. Plots of (■) τ_{λ_0} (lag), (▼) Δt_{λ_0} (picoseconds), and (○) $\tau^* = \tau_{\lambda_0} \times \Delta t_{\lambda_0}$ (picoseconds) as a function of the location of λ_0^{opt1} are shown.

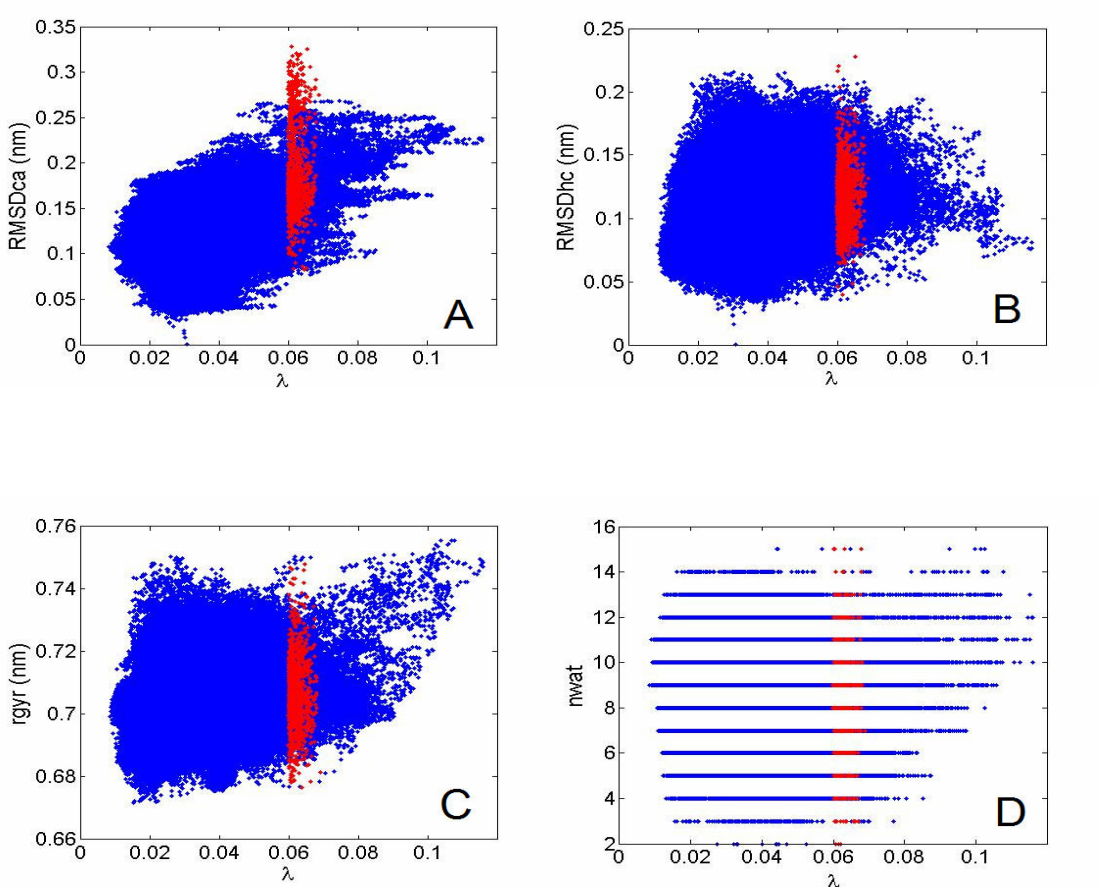


Figure 3.3. Distribution of conformations at $RMSD_{hx} = \lambda_0$ (red) used as starting states for the FFS runs, juxtaposed over the phase space sampling of basin A (native state at 300K) obtained from a 20 ns REX simulation (blue). The plots correspond to $\lambda = RMSD_{hx}$ vs. $RMSD_{ca}$ (A) , $RMSD_{hc}$ (B), r_{gyr} (C) and n_{wat} (D).

Given the significant statistical uncertainty associated with estimating k_{NL}, our FFS study and the TIS work of Juraszek and Bolhuis[18] complement each other by providing comparable rate constant estimates that are in the same order of magnitude of the experimental value reported in Ref. 19.

B. Reaction Coordinate Analysis

The p_B history data obtained with the FFS-LSE method from each of our sets of FFS runs was used to screen a set of candidate collective properties (see Sec. IIIB) to identify an optimized order parameter model. For each case, a tentative regression model, including the ten order parameters monitored during our simulations (see Sec. IIIA) and quadratic interaction terms between them was proposed and used to fit the p_B history data.

For our initial set of FFS runs, the analysis of variance for this model indicated that the terms for $RMSD_{hx}$ and $RMSD_{ca}$ are the only significant ones, yielding the model:

$$\lambda^{opt1}\left(RMSD_{hx}, RMSD_{ca}\right) = p_B = -1.31 + 6.11(RMSD_{hx}) + 3.49(RMSD_{ca}) - 3.0(RMSD_{hx})(RMSD_{ca}) \quad (22)$$

where the $RMSD_{hx}$ and $RMSD_{ca}$ are given in nanometers. The p_B surface predicted by this reaction coordinate model [i.e., Eq. (22)] is illustrated in Fig. 3.4A (red dotted lines), together with the model predicted by Juraszek and Bolhuis (solid lines)[18]; these models have been plotted over the density map of our first set of FFS runs. The isocommittor surfaces of both estimates have similar slopes, with our TSE isoline located at somewhat lower values of $RMSD_{hx}$ and having a very slight curvature. The correlation factor of the model is a rather modest $R^2 = 0.76$.

It is generally expected that:

$$p_B(x) = \begin{cases} \lambda^{opt}\left(RMSD_{hx}, RMSD_{ca}\right) & \mathbf{0} < \lambda^{opt} < 1.0 \\ 0 & \lambda^{opt} \le 0 \\ 1.0 & \lambda^{opt} \ge 1.0 \end{cases} \tag{23}$$

The validity of the $\lambda^{opt1} \le 0$ condition was probed by using the MD data of the $\Phi_{A,0}$ flux calculation to identify the region of attraction of basin A based on the λ^{opt1} model. Figure 3.5A shows a plot of $RMSD_{hx}$ vs. λ^{opt1}, where sampling is observed for $-1 \le \lambda^{opt1} \le -0.4$, with no visits to $\lambda^{opt1} \ge -0.2$. This indicates that trajectories starting at $-0.2 \le \lambda^{opt1} \le 0$ may not always be attracted to basin A and therefore have a non-negligible probability of reaching basin B, contradicting the second condition in Eq. (23). This inconsistency arises in part because: (i) the FFS runs did not sample well states with p_B values close to zero, a reflection of the original choice for λ, and (ii) the λ^{opt1} model obtained has low quality, specially for small values of p_B, a reflection of item (i) and the inherent statistical errors associated with the FFS-LSE approach. As noted by Eq. (23), the λ^{opt1} model should not be used for studying the $\lambda^{opt1} \le 0$ and $\lambda^{opt1} \ge 1$ regions, which lie beyond the region where data were collected and fitted. However, we extrapolated the model for $\lambda^{opt1} \le 0$ as a guide only, to try to identify an alternate λ_0 interface with the method described in Sec. II-B for the new set of FFS runs.

Despite its limitations, the λ^{opt1} model did facilitate the sampling of successful pathways between both basins during our second set of FFS runs. These new runs (using CBG) led to an improved estimate of the reaction coordinate:

$$\lambda^{opt2}\left(RMSD_{hx}, RMSD_{ca}\right) = p_B = 0.75 - 10.8(RMSD_{hx}) - 4.3(RMSD_{ca}) + 60.0(RMSD_{hx})(RMSD_{ca}) \tag{24}$$

The higher quality of λ^{opt2} is evidenced by its $R^2 = 0.92$ correlation factor. The behavior of this model at and around basin A shown in Fig. 3.5B (also obtained from the initial MD simulation for calculating the flux $\overline{\Phi}_{A,0}$), contrasts the one seen in Fig. 3.5A for λ^{opt1}. In Fig. 3.5B, sampling is observed for $-0.1 \leq \lambda^{opt2} \leq 0.5$, corresponding to basin A and its region of attraction and before the expected transition state region.

Figure 3.4B shows the isocommittor surfaces corresponding to the λ^{opt2} reaction coordinate model, Eq. (24), plotted over the density map of our second set of FFS runs. For reference purposes, we have included approximate boundaries for states A (native) and B (loop), extracted from simulations performed at each of these basins. Unlike our λ^{opt1} model (see Fig. 3.4A), it is apparent that the interaction term in λ^{opt2} describes a strong curvature of the p_B isocommittor surface. Model λ^{opt2} indicates that only $RMSD_{hx}$ and $RMSD_{ca}$ are needed to satisfactorily describe the system's progression to state B, and that they change concertedly (cross term) along the most probable transition pathways. A structure with $\lambda^{opt2} = 0.5$ (TS) is included in Fig. 3.6 together with an initial [$\lambda^{opt2} = 0$] and a typical final [$\lambda^{opt2} = 1$] conformation. As observed, the protein's helicity has already been lost in our TS structure. Note also that the representative structure at λ_B^{opt2} is analogous to the one depicting the loop (L) conformation in Fig. 1 of Ref. 18. Further analysis of the TPE is given in the next section.

Pertaining to the TPE sampled during our FFS runs, Fig. 3.4B shows three distinct main pathways leading to basin B. The most visited one corresponds to trajectories crossing through intermediate values of $RMSD_{ca}$, pathway that follows the turning points of the λ^{opt2} isocommittor lines. Two less probable pathways are also observed, comprising trajectories that sample either low or high $RMSD_{ca}$ regions. Conversely, Fig. 3.4A shows limited sampling of the TPE, with most of the successful trajectories following the intermediate "lower energy" pathway of Fig. 3.4B. The

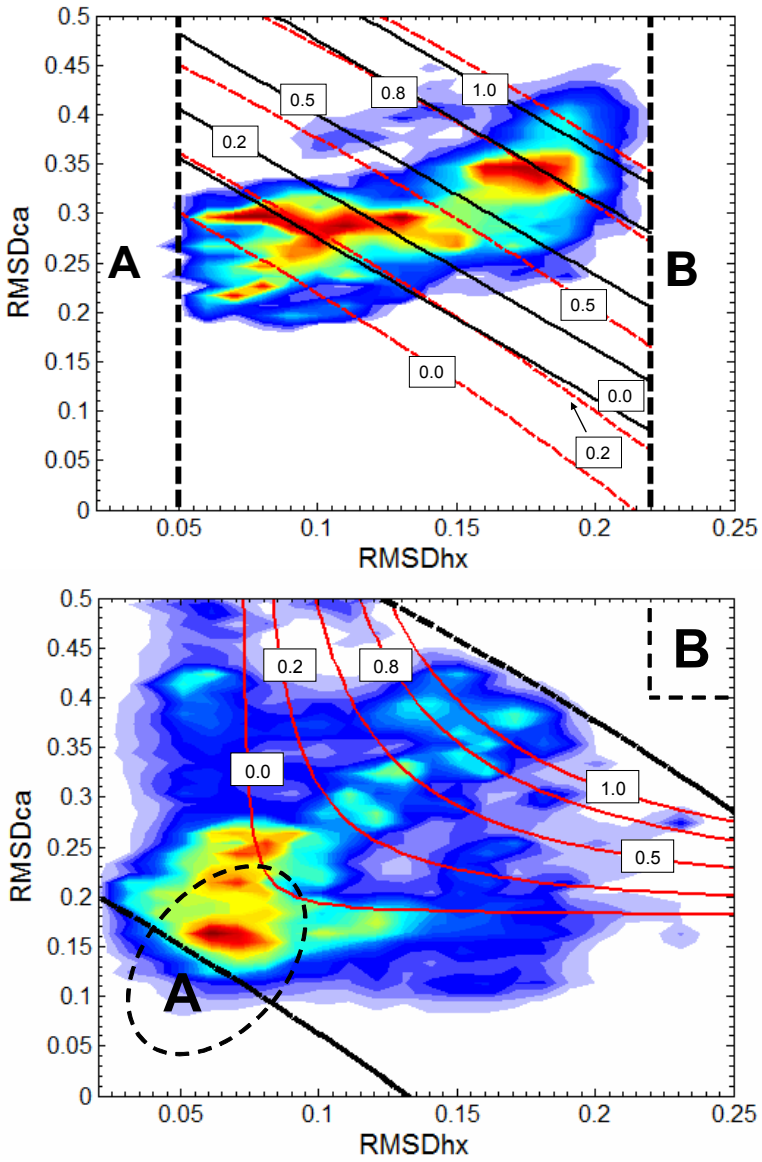


Figure 3.4. Isocommittor surfaces of the reaction coordinate model for the N-L transition: from λ^{opt1} (red dashed line) and Juraszek and Bolhuis[18] (continuous black line) (*A*), and λ^{opt2} (*B*). The isocommittor surfaces are projected over the TPE density maps for the first (*A*) and second (*B*) set of FFS runs, in the region between the two stable states A (native state) and B (loop state). Coloring of the states ranges from least visited (blue/gray) to most visited (red/black). Approximate boundaries for basins A and B are depicted in Figure *B* (bright green).

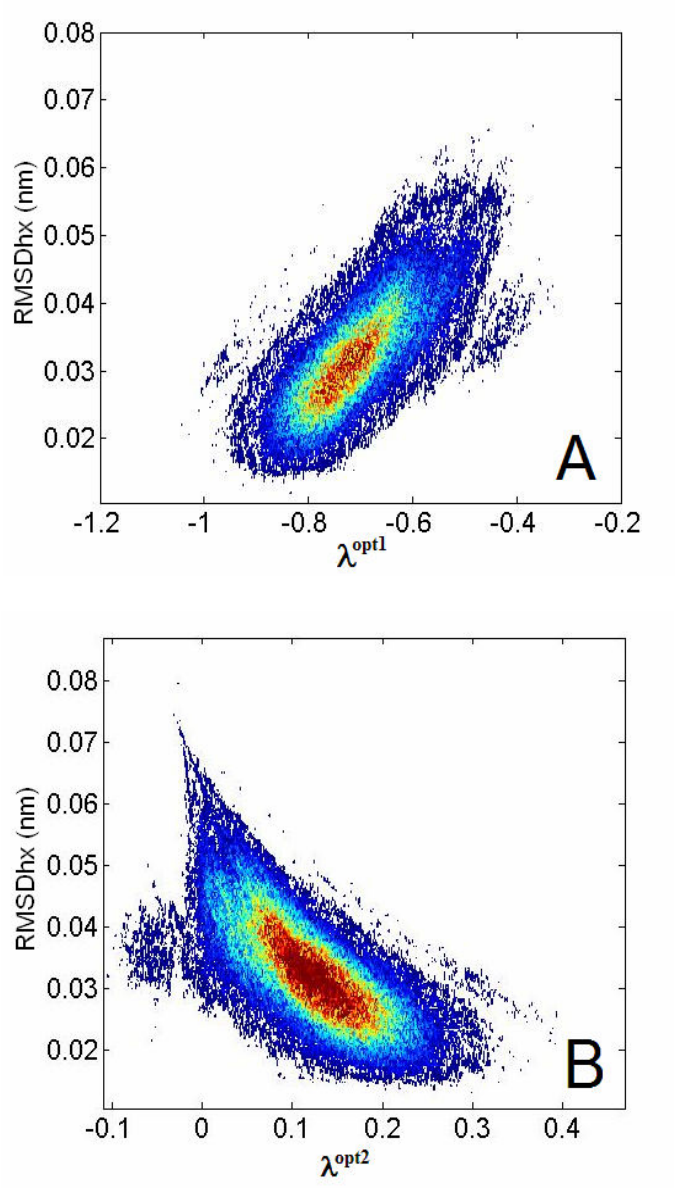


Figure 3.5. Plots of $RMSD_{hx}$ vs. λ^{opt1} (*A*) and λ^{opt2} (*B*) in the region of attraction of basin A, obtained from the MD simulation for the flux $\Phi_{A,0}$ calculation.

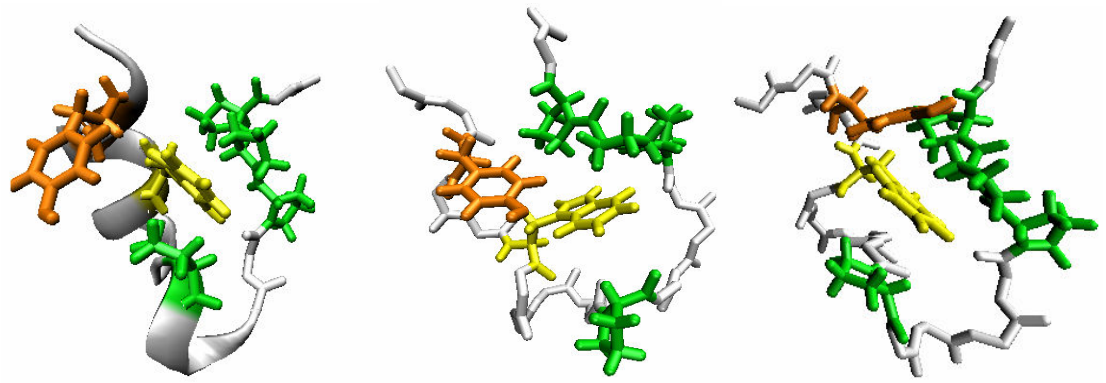


Figure 3.6. Representative structures of conformations at λ_A^{opt2} (left), TSE (middle) and λ_B^{opt2} (right).

improved sampling observed in the second set of FFS runs must be related to the use of a better guiding order parameter (λ^{opt1} vs. $RMSD_{hx}$) and the use of CBG.
A more accurate estimate of the reaction coordinate model could be obtained by performing additional FFS runs. Ideally, these runs should adopt model λ^{opt2} to define the λ interfaces (i.e.; an iteration procedure alluded to in Sec. II C). However, based on our order parameter analysis along the TPE (see Sec. IVC) and the high correlation factor of λ^{opt2}, we surmise that a subsequent iteration would not lead to a considerable improvement of our model.

C. Mechanism

To further elucidate key features of the N-L unfolding mechanism, we now examine how changes in the optimized reaction coordinate λ^{opt2} relate to changes of different order parameters.

Figure 3.7 shows the progression of λ^{opt2} against various order parameters in the region of validity for our p_B estimate ($0 < \lambda^{opt2} < 1$). As expected, the plots for $RMSD_{hx}$ (Fig. 3.7A) and $RMSD_{ca}$ (Fig. 3.7B) show that their higher density regions increase monotonically with λ^{opt2}. Moreover, these two order parameters present linear fits (not shown) with considerably higher correlation factors ($R^2 = 0.41$, for both cases) than the remaining order parameters.

The fraction of native contacts ρ (Fig. 3.7C) has its higher density region decreasing steadily along λ^{opt2}. ρ has the third best correlation factor, $R^2 = 0.18$, but it is less than half the value for $RMSD_{hx}$ and $RMSD_{ca}$, likely due to the broad dispersion of points in the low λ^{opt2} region (R^2 increases to 0.35 when the linear fit is restricted to the regions of higher occurrence in ρ space). This ability to correlate the transition progression is distinctive of many proteins and concurs with the widespread use of

ρ as an acceptable reaction coordinate. In fact, ρ was originally used by Juraszek and Bolhuis[3] to describe the complete Trp-cage N-U transition; our results indicate that while it is a useful guide in the range ($0 < \rho < 0.7$), it has a limited ability to capture later changes. ρ may be an meaningful contributor to the reaction coordinate model when studying the "fast" L-U transition, in which the N-terminal helical content has been lost and $RMSD_{hx}$ is not suitable for capturing the key structural changes that take place (e.g., see Fig. 1 in Ref. 18).

The high correlation factors for $RMSD_{hx}$, $RMSD_{ca}$ and ρ are in clear contrast with the $R^2 < 0.04$ values for the remaining order parameters, buttressing their unsuitability for describing the overall N-L transition. Despite the latter, $RMSD_{hc}$, n_{wat} and the $sb_{(1,2,3)}$ order parameters were found to give insights on likely structural changes occurring along the TPE.

Figure 3.7D shows the progression of $RMSD_{hc}$ along the transition path, where the trajectories appear to have no clear preference on the states visited. Nevertheless, a closer look at the individual pathways showed that many of them tend to visit higher $RMSD_{hc}$ regions ($RMSD_{hc} > 0.2$) at some instance in the range $0.2 < \lambda^{opt2} < 0.7$, suggesting a possible characteristic feature of the N-L pathway. Visual inspection of these high $RMSD_{hc}$ structures indicates that the partial core opening takes place when Pro-12 adopts a position similar to that observed in the TSE structure of Fig. 3.6, farther away from Trp-6 (a propensity also observed in the $p_B \approx 0.5$ structure of Fig. 5-c, Ref. 18), and different from the typical location seen in lower $RMSD_{hc}$ structures such as λ_A and λ_B. This observation is consistent with the work by Piana et al.[6] who suggested that an initial metastable state of the unfolding pathway may be characterized by a partial opening of the hydrophobic core. Interestingly, Juraszek and Bolhuis[3] have reported the presence of a P_d state along their alternate unfolding route, also identified by a detachment of Pro-12 from the hydrophobic core. This behavior is

consistent with our findings, and may be related to the N-P_d-L route proposed in Refs. 3, 18.

The progression of sb_2 along our reaction coordinate is shown in Figure 3.7E (sb_1 and sb_3 display an analogous behavior). Two distinct regions are consistently sampled along the TPE, one within the range $0.5 < sb_2 < 0.7$ and the other one constrained to $0.15 < sb_2 < 0.21$. During the initial stages of the transition ($\lambda^{opt2} < 0.4$), the lower sb_2 distance is visited most frequently, whereas the higher sb_2 distance is predominantly sampled after the transition state. Fewer low-sb_2 distance occurrences (i.e., less structures with a tight Asp-9 - Arg-16 salt bridge) as λ^{opt2} evolves is consistent with previous studies that have proposed that the presence of Asp-9 - Arg-16 salt bridges stabilizes the folded state.[8,31]

Finally, regarding the number of water molecules around Trp-6, no clear solvation pattern is apparent along λ^{opt2} (see Fig. 3.7F), indicating the absence of a distinct effect of water dynamics in the reaction coordinate. However, an increase in the number of structures with a heavily solvated Trp-6 (i.e., $n_{wat} > 18$) was observed for $\lambda^{opt2} > 0.5$; this implies a higher propensity for adopting an open loop conformation. The wide variety of Trp-6 solvated states achievable at each interface (which is observed even within basin A, see Fig. 3.3D) is in agreement with the TPS results by Juraszek & Bolhuis[18] (see Fig. 4-a of Ref. 18), and suggests that explicit solvation may be required for proper modeling of this protein. Further studies in which the location of water molecules around Trp-6 is mapped out, may help elucidate specific water-protein interactions at play in the unfolding process.

Overall, the N-L unfolding mechanism observed in our simulations is fairly consistent with the one proposed by Juraszek and Bolhuis,[3,18] characterized by a steady preservation of the U-shaped structure, and comprised of three main stages: i) an early ($\lambda^{opt2} < 0.1$) destabilization of the 3_{10}-helix, ii) a progressive loss of helicity

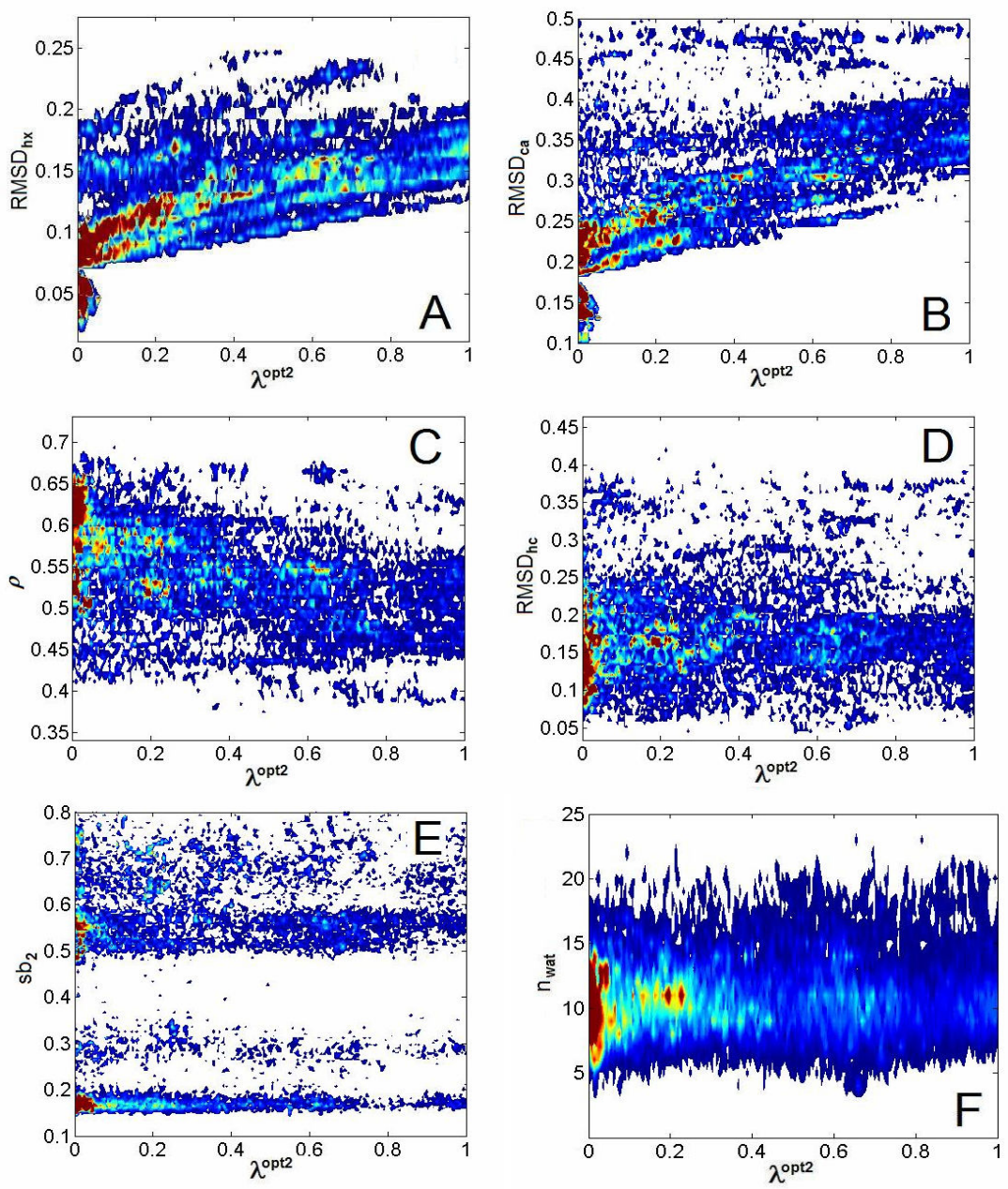


Figure 3.7. Optimized reaction coordinate model vs. various order parameters along the TPE. The plots correspond to λ^{opt2} vs. $RMSD_{hx}$ (*A*), $RMSD_{ca}$ (*B*), ρ (*C*), $RMSD_{hc}$ (*D*), sb_2 (E) and n_{wat} (*F*).

with virtually complete destabilization of the α-helix before the transition state (point where changes in $RMSD_{ca}$ become more important), and iii) a transitory increase in the tendency towards a partially open hydrophobic core, induced by a detachment of Pro-12 from Trp-6, which likely promotes a higher solvation of Trp-6 in the final stages of the transition. In this latter stage a greater disruption of the Asp-9 - Arg-16 salt bridges may also occur leading up to the loop state. Qiu & Hagen[32] have highlighted the significance of enthalpic contributions for the preservation of the native conformation. In this sense, pathways are likely to "speed up" their approach towards basin B once the key electrostatic interactions that hold the native structure together are weakened or lost. In future studies, a more detailed mechanism of the N-L transition could be obtained by analyzing alternate interactions (other than the Asp-9 - Arg-16 salt bridge).

V. FINAL REMARKS

Our results validate the use of optimized FFS-MD methods as appropriate schemes for elucidating the kinetics and mechanism of the Trp-cage unfolding N-L transition in explicit solvent. By applying novel FFS algorithms[22-23] we were able to overcome some of the potential limitations that arise while implementing traditional FFS methods.[16-17] For the case studied, ample sampling around basin A leading to suitable selection of the λ_0 ensemble appears to be critical for successful implementation of the method. Our protocol to obtain states at the initial interface provides a simple way to ensure an uncorrelated, properly sampled λ_0 ensemble from FFS simulations in complex systems. More generally, our work suggests that the successful implementation of novel path sampling methods (and of FFS in particular) crucially depends on the careful selection of the method's parameters; this selection should not be based only on experience but also on systematic optimization strategies

such as those used in this work. The agreement between the rate constant values found using $\lambda = RMSD_{hx}$ and $\lambda = \lambda^{opt1}$ suggests that consistent estimates for the rate can be readily obtained from FFS with an "unoptimized" λ, provided that it sufficiently correlates the overall changes through the transition (e.g., $RMSD_{hx}$ or $RMSD_{ca}$ in this study).

The N-L transition was found to follow uncorrelated pathways along most of the order parameters analyzed, except for $RMSD_{hx}$ and $RMSD_{ca}$, whose evolution shows systematic changes between the native and the loop state. The reaction coordinate model λ^{opt2} obtained from our FFS-LSE analysis corroborates the significance of these two properties for capturing the key changes during the N-L transition. The monotonic increase of λ^{opt2} as a function of $RMSD_{hx}$ and $RMSD_{ca}$ contrasts the behavior of the remaining order parameters tested (see Fig. 3.7). Other events such as partial disruption of the hydrophobic core, disruption of the Asp-9 - Arg-16 salt bridge, and increased solvation of Trp-6 seem to be meaningful around or after the TSE, as markers of particular but non-sequential events that characterize the N-L route.

Ongoing studies are aimed at applying optimized FFS methods to other systems of interest (e.g., other biomolecular transitions in the μs timescale, and processes whose study may have posed difficulties to other path sampling methods). In addition, we are currently evaluating other techniques for enhancing the efficiency of FFS, for instance, by harnessing accelerated dynamic methods and Markovian models. Such new algorithms are needed to study biological systems exhibiting multiple intermediates and transition channels.

ACKNOWLEDGMENTS

The authors are grateful for support from the National Science Foundation, Award No. CBET-0933092.

REFERENCES

1. J. L. Alonso and P. Echenique, Biophys. Chem. 115, 159 (2005).

2. Z. Hu, Y. Tang, H. Wang, X. Zhang, and M. Lei, Arch. Biochem. Biophys. 475, 140 (2008).

3. J. Juraszek and P. G. Bolhuis, Proc. Natl. Acad. Sci. U. S. A. 103, 15859 (2006).

4. D. Paschek, S. Hempel, and A. E. Garcia, Proc. Natl. Acad. Sci. U. S. A. 105, 17754 (2008).

5. D. Paschek, H. Nymeyer, and A. E. Garcia, J. Struct. Biol. 157, 524 (2007).

6. S. Piana, and A. Laio, J. Phys. Chem. B 111, 4553 (2007).

7. L. Yang, M. P. Grubb, and Y. Q. Gao, J. Chem. Phys. 126, 125102 (2007).

8. R. Zhou, Proc. Natl. Acad. Sci. U. S. A. 100, 13280 (2003).

9. A. Linhananta, J. Boer, and I. MacKay, J. Chem. Phys. 122 (2005).

10. A. Schug, T. Herges, A. Verma, K. H. Lee, and W. Wenzel, Chem. Phys. Chem. 6, 2640 (2005).

11. A. Schug, W. Wenzel, and U. H. Hansmann, J. Chem. Phys. 122, 194711 (2005).

12. P. J. Steinbach, Proteins 57, 665 (2004).

13. B. Zagrovic, and V. S. Pande, Nat. Struct. Biol. 10, 955 (2003).

14. C. Dellago, P. G. Bolhuis, F. S. Csajka, and D. Chandler, J. Chem. Phys. 108, 1964 (1998).

15. D. Moroni, T. S. van Erp, and P. G. Bolhuis, PHYSICA A 340, 395 (2004).

16. R. J. Allen, D. Frenkel, and P. R. ten Wolde, J. Chem. Phys. 124, 024102 (2006).

17. R. J. Allen, P. B. Warren, and P. R. ten Wolde, Phys. Rev. Lett. 94, 018104 (2005).

18. J. Juraszek and P. G. Bolhuis, Biophysical Journal 95, 4246 (2008).

19. L. Qiu, S. A. Pabit, A. E. Roitberg, and S. J. Hagen, J. Am. Chem. Soc. 124, 12952 (2002).

20. C. Dellago and P. G. Bolhuis, Transition path sampling and other advanced simulation techniques for rare events. In *Advanced Computer Simulation Approaches for Soft Matter Sciences III*, Eds. C. Holmer, K. Kremer: Springer, 2008; Vol. 221, pp 167.

21. F. A. Escobedo, E. E. Borrero, and J. C. Araque, J. Phys.: Condens. Matter 21, 333101 (2009).

22. E. E. Borrero and F. A. Escobedo, J. Chem. Phys. 127, 164101 (2007).

23. E. E. Borrero, and F. A. Escobedo, J. Chem. Phys. 129, 024115 (2008).

24. C. Velez-Vega, E. E. Borrero, and F. A. Escobedo, J. Chem. Phys. 130, 225101 (2009).

25. R. J. Allen, D. Frenkel, and P. R. ten Wolde, J. Chem. Phys. 124, 194111 (2006).

26. D. Moroni, T. S. van Erp, and P. G. Bolhuis, PHYSICA A 340, 395 (2004).

27. A. Ma and A. R. Dinner, J. Phys. Chem. B 109, 6769 (2005).

28. D. Van Der Spoel, E. Lindahl, B. Hess, G. Groenhof, A. E. Mark, and H. J. Berendsen, J. Comput. Chem. 26, 1701 (2005).

29. Y. Sugita, Y. Takahashi, I. Hayashi, M. Morimatsu, K. Okamoto, and M. Shigemori, Pathol. Int. 49, 1114 (1999).

30. P. Brenner, C. R. Sweet, D. VonHandorf, and J. A. Izaguirre, J. Chem. Phys. 126, 074103 (2007).

31. J. W. Neidigh, R. M. Fesinmeyer, and N. H. Andersen, Nat. Struct. Biol. 9, 425 (2002).

32. L. Qiu, and S. J. Hagen, J. Am. Chem. Soc. 126, 3398 (2004).

33. E. E. Borrero and F. A. Escobedo, J. Phys. Chem. B 113, 6434 (2009).

34. J. Rogal, and P. G. Bolhuis, J. Chem. Phys. 129, 224107 (2008).

CHAPTER 4

CORRELATING STRUCTURAL AND SOLUBILITY BEHAVIOR OF SELECTED Aβ-42 POLYPEPTIDE MUTANTS

I. INTRODUCTION

Alzheimer's disease (AD) is a major health issue worldwide. In the US alone, more than 5 million people are affected, a number that is projected to triple by the year 2030. AD is a neurodegenerative condition, pathologically characterized by the accumulation of extracellular plaques of Amyloid β-protein (Aβ), and the intracellular formation of neurofibrillary tangles of tau β-protein. The Aβ cascade hypothesis is currently the dominant mechanism for the onset and progression of AD, a process believed to be triggered by an age-related increase in the production of the Aβ-42 protein (amyloid peptides of length 42 aminoacids) relative to the more common and less neurotoxic Aβ-40 peptide.[1] An increase in the Aβ-42/Aβ-40 ratio promotes aggregation of improperly folded Aβ monomers, leading to the formation of oligomers and amyloid plaques/fibrils, and ultimately to neuron cell damage. In the past decade, an increasing number of studies have reported that oligomers, and not fibrils, may be the primary neurotoxic agents (for a review, see ref. [1]). As a result, much of the research has thereafter shifted from the study of fibril formation pathways toward the elucidation of monomer/oligomer structural characteristics and their aggregation mechanisms. Low molecular weight oligomers ranging from dimeric to octameric aggregates are currently believed to be the smallest soluble Aβ species responsible for decreased synapse density, a marker that best correlates with the extent of dementia in AD.[2-6] Further examination of the Aβ-40 and Aβ-42 aggregates (which appear to follow different oligomerization pathways[3,6]) has led to a common belief that the

structure of the oligomers varies with size and monomer type.[3,5,7] Unfortunately, experimental studies aimed at the detailed characterization of oligomeric and monomeric structures (*e.g.*, crystallization) at physiological conditions have been greatly hindered by the Aβs' high aggregation rates, as well as their sensitivity to specific physicochemical conditions. Nevertheless, distinctive features of the monomers in water have been elucidated using NMR techniques for the Aβ-40[8] and Aβ-42[9] peptides. Other studies have focused on structural resolution via less amyloidogenic Aβ fragments,[8,10-12] or within environments that discourage aggregation.[13-17]

Further insights on the Aβ monomer/oligomer structures and their aggregation mechanism have been derived from diverse computational approaches. Numerous groups have studied specific segments believed to be central for aggregation or folding nucleation (using implicit[18-23] or explicit[24-33] solvent), whereas others have modeled the complete Aβ wildtype/mutant structures (using implicit[34,35] or explicit[36-39] solvent). Notably, dissimilar results for analogous systems are found in many of these studies; such inconsistencies can be mainly attributed to the specific sequence and length of the modeled segments, as well as to the effect of the force field and solvation model on the dynamics of this flexible peptide. While most of these studies may correctly describe monomer/oligomer features, like total β-sheet/helical content, that have been estimated from experimental analyses (which have also reported inconsistent findings), full validation of an appropriate *in silico* model for Aβs and their aggregates at physiological conditions would entail comparison with experimental structures, which are still unknown. Despite the latter, researchers have been able to improve the reliability of their simulations by matching their models with available experimental data (e.g., NMR restraints), or by comparison of structures from different force fields and solvation models. Of particular interest are the studies by Sgourakis et al.[36] and

Krone et al.,[29] who independently identified the OPLS-AA[40]/TIP3 solvent[41] model as one that appropriately represents the 21-30 fragment[29] and the full Aβ monomers.[36] These results are consistent with our own preliminary search for a satisfactory Aβ model, in which we explored different force fields with implicit and explicit solvent models. Moreover, we found that the study of the complete monomer in explicit solvent is important for the detection of structural features (e.g., secondary structure content at the Central Hydrophobic Core,[9] RES. 17-21) and specific electrostatic interactions (e.g., those between E22 or D23 and K28[42]) observed in experiments.

In the present study, we report our findings on the structural features of Aβ-42 that may promote dimerization (as a first step towards higher order oligomerization), by performing simulated analyses on the complete wildtype (WT) Aβ-42 and two mutants of this peptide. We carried out explicit solvent simulations using a novel adaptation of Replica Exchange Molecular Dynamics (REM),[43] called All Pairs Exchange (APE),[44] that significantly enhances the efficiency of REM sampling. The Aβ-42 variants analyzed were selected such that they displayed either a notably lower (soluble variant) or higher (insoluble variant) aggregation rate with respect to the one observed for WT Aβ-42. They were also required to have the fewest possible number of mutations, since this reduces the complexity of identifying key differences that may lead to dissimilar aggregation rates.

We chose GM6 (F19→S19, L34→P34) as the soluble variant, a peptide that has consistently displayed virtually no aggregation in several *in vitro* studies[45-47] using different "folding quality" assays. To our knowledge, a structural analysis of the full GM6 variant has not yet been performed and is therefore a very attractive choice for this study. We selected Dutch (E22→Q22) Aβ-42 as the *insoluble* variant, given that it has shown a considerably faster *in vitro* aggregation rate than that of WT Aβ-42.[48] Most of the studies on the Dutch type have focused on the 40 residue monomer, due to

its apparent role in hereditary cerebral hemorrhage with amyloidosis Dutch-type (HCHWA-D).[49-52] However, we targeted our analysis on the 42 length monomer, given that both its *in vitro* neurotoxicity and aggregation rate are appreciably higher than those of the WT Aβ-42 and Dutch Aβ-40 variant;[48,53] in addition, this selection allows the study of relevant discrepancies between electrostatic interactions involving residues 41 and 42.* The Dutch variant has been studied using different segments (e.g., RES 15-28,[24] RES 10-35[54], RES 21-30[55]) . However, to our knowledge, no detailed structural model on the complete monomer has yet been reported.

Sec. II provides a brief description of our model setup and simulation approach. In Sec. III we present the results for the Aβ-42 mutagenesis analysis. Finally in Sec. IV we discuss our findings and give some concluding remarks.

II. METHODS

Monomer simulations

The configuration space of the Dutch, WT, and GM6 Aβ-42 monomers was explored using the OPLS-AA/TIP3P water model, via REM/APE simulations in GROMACS[56] molecular simulation package. The APE method considerably increases the probability of generating an exchange between pairs of replicas, while meeting the detailed balance condition. For various systems previously studied by our group, REM/APE reduced at least by a factor of two the simulation time required for configurational sampling compared to conventional REM.[57]

The systems were prepared as follows. The structure of Aβ-42 in an apolar solvent (PDB code: 1IYT)[14] was mutated for the Dutch and GM6 variants and energy minimized using the Steepest Descent algorithm. After peptide solvation and

* Unless otherwise noted, for the remainder of this document it is assumed that all mutant types are 42 aminoacids long.

neutralization of the system, we heated the resulting structure to 700 K and carried out a 10 ns MD simulation at constant temperature using Nose-Hoover[58,59] thermostat (employed for all of our simulations), from which a random coil structure was obtained[**]. The time step for all our simulations was 2 fs, permitted by the use of LINCS[60] algorithm for constraining bond lengths. The coiled conformation was then collapsed by means of a 5 ns vacuum simulation at the same 700 K that allowed resolvation of the peptides in 3393 (Dutch=3371, GM6=3401) molecules of TIP3P water. This was followed by a short MD run at 300 K for equilibration of the water box, in which the position of the peptide was restrained. A 1 ns MD simulation at P = 1 atm and T = 300 K was then carried out for equilibration of the whole water-protein system. The WT, Dutch and GM6 structures thus obtained were used as starting conformations for our REM/APE simulations. We note that our procedure for generating initial structures, analogous to the one used by Sgourakis et al.,[36] considerably facilitates the simulation of the complete Aβ structure in explicit solvent by solvating a rather collapsed peptide instead of an otherwise extended conformation, which would require a significantly higher number of water molecules for its solvation. It could be argued that sampling may be hampered when a collapsed coil structure is used as initial conformation. However, a detailed validation of this model using NMR ^{3}J-coupling constants[36] and the use of REM/APE to promote rapid conformational sampling, makes this approach very suitable for modeling the dynamics of Aβ-42.

REM/APE simulations were carried out for the three cases in the 250-600 K range. Swaps were attempted every 1 ps and an exchange probability close to 20% was targeted, requiring 32 replicas that were exponentially distributed along the

[**] The solvated random coil structures were used for exploratory studies aimed at optimizing the simulation time required for this analysis.

temperature range. All of the systems were run for 25 ns/replica (a total of 0.8 μs for each simulation), and configurations were saved every 1 ps. For each case, the data acquired for the replica at 296 K was used for the analysis presented in the results section. We note that, in all cases, the statistics obtained for three other replicas at 288, 305 and 313 K are analogous to the ones reported at 296 K. All of our REM/APE simulations were run in the NIC of Corning Inc.

Analysis Tools

Unless otherwise noted, our analyses are performed on the ensemble gathered at room temperature (i.e., 296 K and 298 K for our REM/APE and MD simulations, respectively). The Single Linkage and Jarvis Patrick clustering methods available through the g_cluster tool in GROMACS were employed to group the 10,000 conformations analyzed for each monomer. Both methods identified analogous dominant clusters for all cases; however, the more stringent Jarvis Patrick algorithm consistently produced a higher number of clusters. Given that for the three monomers studied a few clusters are able to group the majority of the ensemble structures, the results presented in the *cluster analysis* section correspond to those obtained via the Single Linkage method. The central (i.e., representative) structures for each cluster were obtained directly from these calculations.[56] Contact maps were generated using the g_mdmat tool in GROMACS, which identifies the minimum distance between residues by calculating the smallest distance between any pair of atoms belonging to distinct residues. A truncation distance of 1.5 nm was employed for the contact map calculations. For analysis of hydrogen bonds we used a cutoff distance of 3.5 A.

Schematic representations of peptides were achieved via either VMD[61] or SwissPDB[62] programs. All other calculations were performed using tools available in GROMACS.

III. RESULTS

In general, the presence of β-sheets/β-bridges and helices in our simulation ensembles agrees qualitatively with experimental studies that have observed secondary structure content in the Dutch,[48] WT[10,17] and GM6[45] monomers. Quantitative agreement is not sought, given the qualitative nature of the experimental assays performed and the high rate of aggregation of the WT and Dutch peptides that may promote conformational changes upon oligomerization in early stages of the experiments. Other features, thought to be distinctive of Aβ monomers, are also observed in our simulations and are thus highlighted in this section. In accordance with our main goal, several analyses were performed so as to identify likely markers of the peptides' dissimilar oligomerization tendencies.

Structural Stability

As an initial assessment of the relative stability of the three peptides, we calculated the backbone Root Mean Square Fluctuation ($RMSF_{back}$, understood as the standard deviation of the backbone's atomic positions with respect to their mean values) of the complete Aβ-42 monomers and their key regions, namely the Central Hydrophobic Core (CHC, RES. 17-21) and the 10-residue N and C terminal segments. Fig. 4.1 shows the average $RMSF_{back}$ (nm) values and corresponding standard deviation (SD) of the regions analyzed, for the ensembles at room temperature (296 K) during the 15-25 ns period. When all residues are considered, a relative decrease in the flexibility of the peptides (i.e., $\overline{RMSF}_{back}^{Dutch} > \overline{RMSF}_{back}^{WT} > \overline{RMSF}_{back}^{GM6}$) is observed but the differences are not statistically significant. Indeed, the magnitude of the backbone motions of these predominantly unstructured monomers is comparable, as evidenced by their high SD values (around 0.15 nm for all cases). In contrast, upon analysis of

the $\overline{\text{RMSF}}_{\text{back}}$ for the key Aβ regions of the three peptides, we found a clear difference in the stability of their N-terminal (RES. 1-10, Fig. 4.1) segments. Interestingly, the relative decrease in the N-terminal stability of the GM6, WT and Dutch variants, correlates with the experimentally observed decline in their relative aggregation rate. We note that, as may be surmised from Fig. 4.1, the C-terminal and CHC regions of all variants are quite flexible and their relative variation in $\overline{\text{RMSF}}_{\text{back}}$ is not statistically significant. Some experimental[63] and computational[36] studies have found the C-terminal of WT Aβ-42 to be less flexible than that of the more soluble WT Aβ-40, leading to the proposal that an increasingly stable C-terminal is more likely to seed aggregation. Our $\overline{\text{RMSF}}_{\text{back}}$ results suggest that both a more soluble (GM6) and a less soluble (Dutch) Aβ-42 variant can have a C-terminal that is just as flexible as that of WT Aβ-42.

The Hydrophobic Solvent Accessible Surface Area ($\text{SASA}_{\text{H}\phi}$) of the system was also monitored as potential marker of the monomers' relative solubility. Fig. 4.2 shows the ensembles' $\text{SASA}_{\text{H}\phi}$ and corresponding SD for the complete peptide and the key Aβ regions, normalized by the number of residues analyzed for each case. Notably, the only statistically significant differences in $\text{SASA}_{\text{H}\phi}$ among the monomers are those observed between the N-terminal and CHC segments of GM6, and the corresponding regions of the WT and Dutch variants. The visibly lower $\text{SASA}_{\text{H}\phi}$ value of GM6's CHC region is a direct consequence of the F19S (nonpolar → polar) mutation in this short region. This is not the case for RES. 1-10, given that this segment is the same for all mutants. A reduced $\text{SASA}_{\text{H}\phi}$ in this N-terminal region of the GM6 mutant implies a lower energetic cost for solvation, in congruence with the higher solubility found experimentally for this mutant.[45] In contrast, comparable $\text{SASA}_{\text{H}\phi}$ values between the C-terminal regions of the Dutch, WT and GM6 peptides

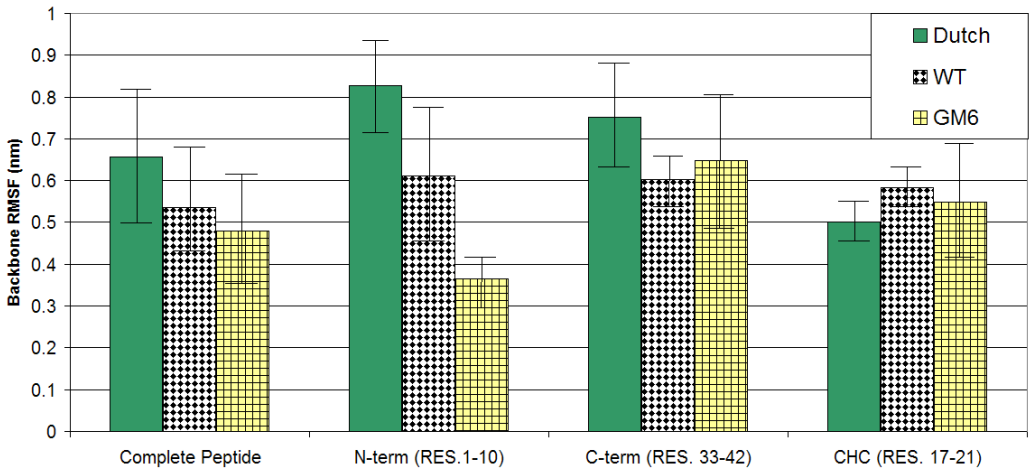


Figure 4.1. Average Backbone RMSF for the monomers studied. Values for the complete protein, N-terminal, C-terminal and CHC regions are shown. The bars represent the standard deviation for each case.

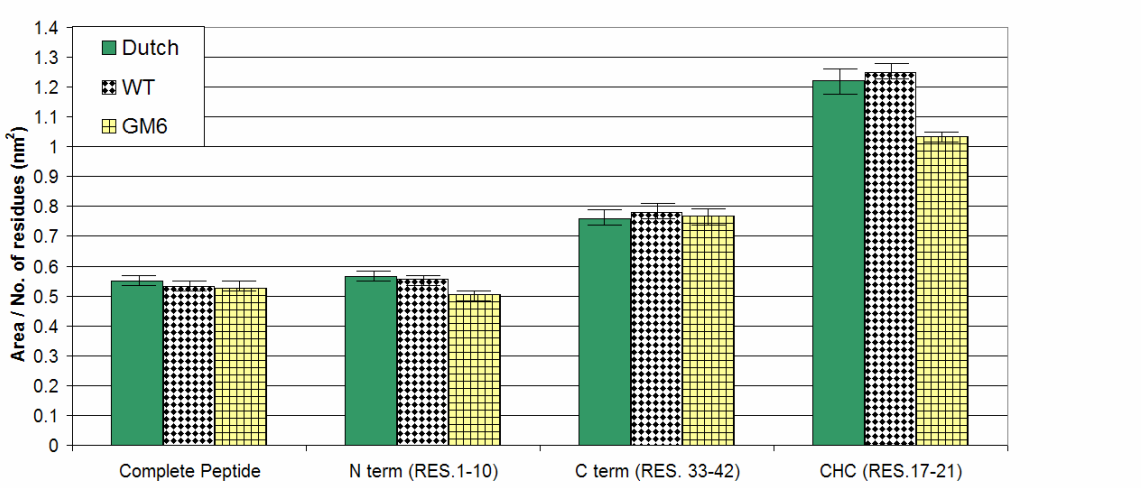


Figure 4.2. Hydrophobic Solvent Accessible Surface Area of the monomers studied, normalized by the number of residues in each segment. Values for the complete protein, N-terminal, C-terminal and CHC regions are shown. The bars represent the standard deviation for each case.

do not support the hypothesis of C-terminal hydrophobic tail clustering as the main driving force for Aβ oligomerization.[34]

Lastly, the total energy of the solvated system (*E*) was measured as a tentative marker of the monomers' relative "folding quality". Specifically, for systems of analogous size, given that the enacted point mutations introduce relatively small perturbations to the basal energy of the system, it is expected that the more "native-like" variants readily sample lower energy regions. Respectively lower energy regions are sampled by the Dutch ($E_{average}$= -108,553 kJ/mol, σ = 682 kJ/mol), WT ($E_{average}$= -114,003 kJ/mol, σ = 522 kJ/mol), and GM6 ($E_{average}$= -118,178 kJ/mol, σ = 452 kJ/mol) variants. These results qualitatively correlate the monomers' "folding quality" with their relative aggregation tendency, and are consistent with the commonly accepted protein misfolding and aggregation hypothesis.[64] More accurate methods that can factor out completely the effect of the mutations (e.g., by comparing only the interaction energies of the common residues, including those of residues with the water molecules) are currently being explored.

Cluster Analysis

In order to further characterize the three monomers under study, we carried out an analysis of the various clusters found within the peptides' ensembles at 296 K, for the 15-25 ns period. The ensembles were grouped into 10 (Dutch), 9 (WT) or 7 (GM6) clusters using an RMSD cutoff of 3 A. Fig. 4.3 shows the cluster population for each case; the central structures of the main clusters, representing about 60% of the ensemble for each case, are shown as insets in each plot. In addition, Fig. 4.4 presents the contact maps for the major clusters of each monomer.

The representative structure of the Dutch variant's dominant cluster (Fig. 4.3A), accounting for 33% of this mutant's ensemble, displays a C-terminal β-hairpin

Figure 4.3. Relative population of the clusters identified for the Dutch (*A*), WT (*B*) and GM6 (*C*) ensembles at 296 K, for the 15-25 ns period. The representative structures of the major clusters for each case are shown as insets in the plots, colored by structure type (yellow=β-sheet, blue=helix, tan=β-bridge, cyan=turn+bend, and white=random coil).

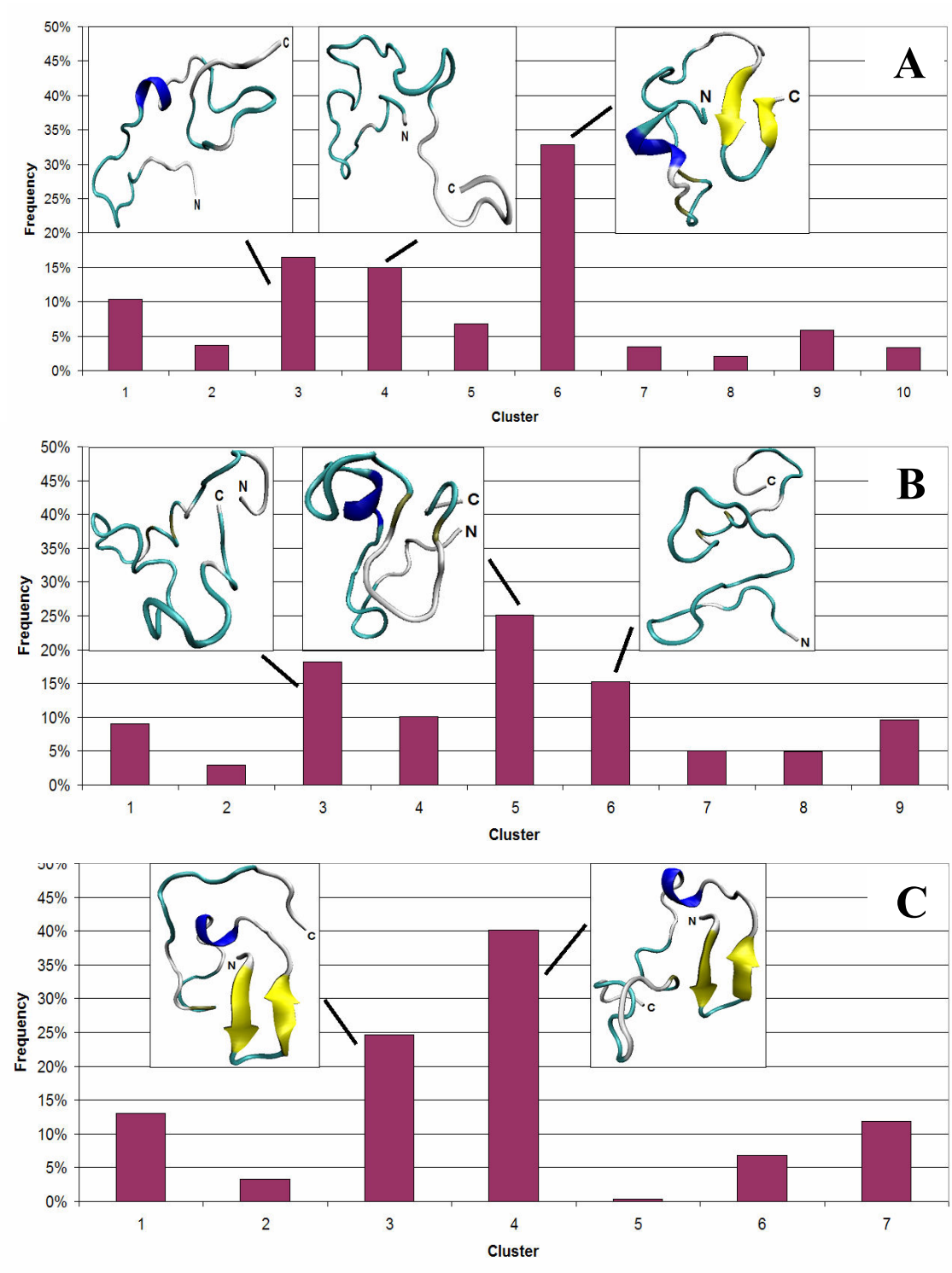
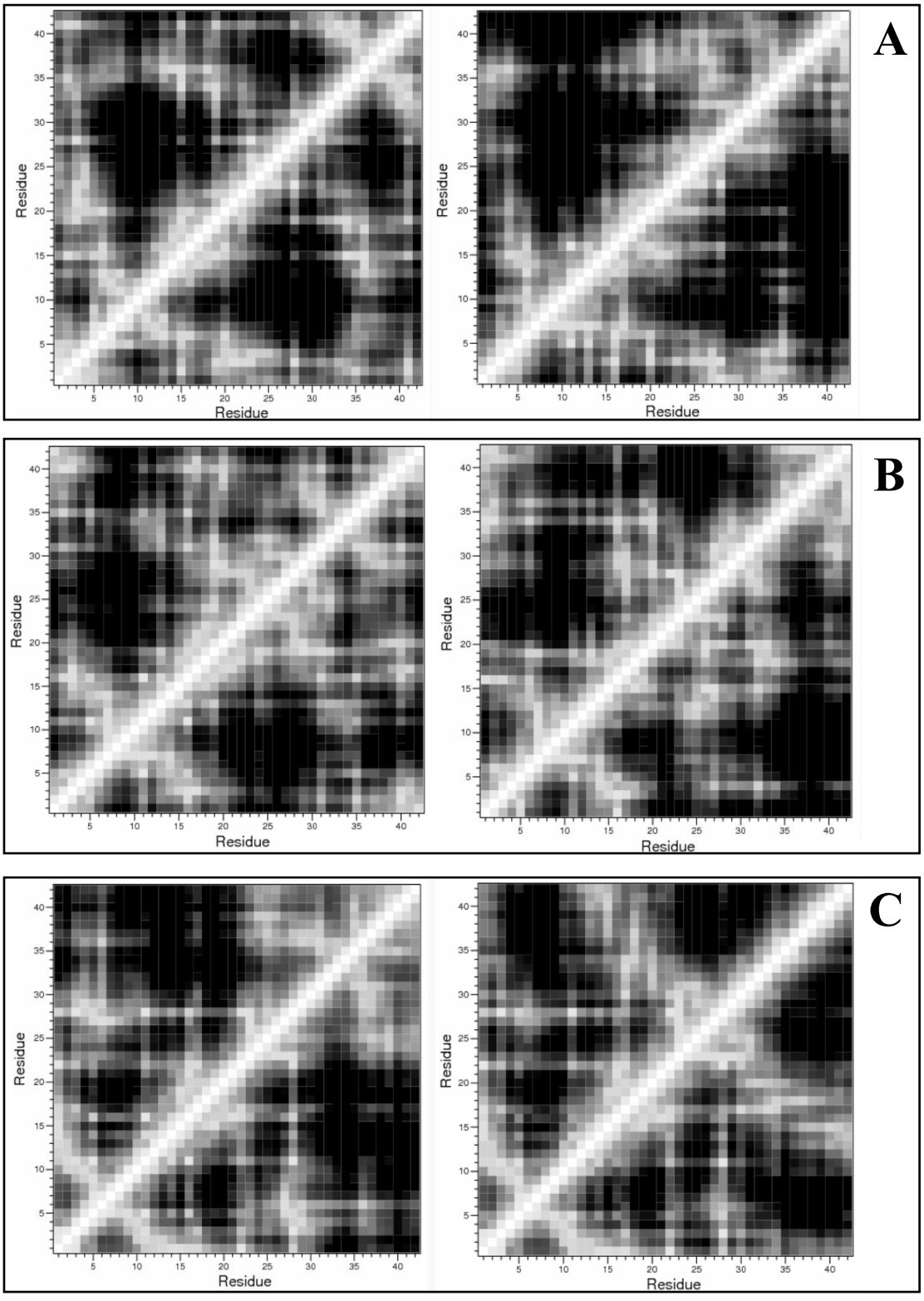

A
Frequency
50%
45%
40%
35%
30%
25%
20%
15%
10%
5%
0%
1
2
3
4
5
6
7
8
9
10
Cluster
C
N
N
C
N
C
B
Frequency
50%
45%
40%
35%
30%
25%
20%
15%
10%
5%
0%
1
2
3
4
5
6
7
8
9
Cluster
C
N
C
N
C
N
C
45%
40%
35%
30%
25%
20%
15%
10%
5%
0%
1
2
3
4
5
6
7
Cluster
N
C
N
C

Figure 4.4. Contact maps for the representative structures of the three major clusters of the Dutch (*A*) and WT (*B*) monomers, and the two major clusters of GM6 (*C*). The clusters presented account for 65% (Dutch, Fig. *A*), 59% (WT, Fig. *B*) and 64% (GM6, Fig. C) of the studied ensemble. The left plots of all figures are symmetric across the diagonal and correspond to the dominant cluster for each case. The right plots of Figs. *A* and *B* are bisected into an upper left (second-most dominant cluster) and a lower right (third-most dominant cluster) section. The right plot of Fig. *C* is again symmetric and corresponds to the second-most dominant cluster for that mutant. The x and y scales for all plots go from 1-42, with ticks representing the residue numbers. The minimum distances between residues are represented by means of a color scale ranging from 0 nm (white) to 1.5 nm (black).


A
B
C
Residue
Residue
5
10
15
20
25
30
35
40

in the ^{33}**GLM**VGG**VVI**^{41} region, a CHC 3_{10} helix and a β-bridge between residues 7 and 12. The cluster's contact map (Fig. 4.4A, left) also shows that the N-terminal tail (RES. 1-5) is able to come in close contact with most of regions of the peptide, evidencing its high N-terminal flexibility. Close proximity between the CHC and C-terminal RES. 33-36 was also evident from this cluster's contact map.

Concerning the remaining two clusters of the Dutch monomer, accounting for 32% of the structures, their central conformations (Fig. 4.3A) lack secondary structure content except for the 3_{10} helix formation at the CHC. Upon scrutiny of their contact maps (Fig. 4.4A right, see figure caption) we see that the N and C terminal tails (RES. 1-5 and RES. 38-42) rarely come in close contact with each other, consistent with the elongated shapes shown in Fig. 4.4A. The contact map and central structure of the second-most dominant cluster suggests a bend in RES. 5-18 that protrudes out of the remaining peptide. Conversely, the third-most dominant cluster shows an elongated C-terminal (RES. 31-42) that rarely comes close to the apparently collapsed rest of the peptide, as well as extensive contacts between the N-terminal tail and the RES. 10-35 segment.

Figs 4.3B and 4.4B illustrate the representative structures and contact maps of the three main WT clusters. The dominant cluster, accounting for 25% of the ensemble, displays a 3_{10} helix at the CHC, a reasonably structured (U shaped) C-terminal stabilized by a β-bridge between residues 30 and 37, and a disordered hairpin motif consisting of two closely interacting strands formed by RES. 3-8 and RES. 12-18. On the other hand, the two other dominant clusters, representing 33% of the population, are mainly unstructured (except for the 3_{10} helix at the CHC). The contact maps for the three clusters show extensive contacts being formed between residues in the 20-30 segment, suggesting that this region is the "hinge" for the WT monomer. In addition, the two dominant clusters show recurrent contacts between the N and C

terminal tails (RES. 1-5 and RES. 38-42). More specifically, we identified conserved H-bonds between D1-A42, A2-V40 and E3-V40 in 65% of the ensemble structures, interactions that likely contribute to the increased stability of the WT Aβ-42 C-terminal region with respect to that of the Dutch and WT Aβ-40 variants.

Interestingly, Sgourakis and co-workers[36] report the presence of a C-terminal β-hairpin (^{31}**IIGL**MVG**GVVI**^{42}) in the dominant (21% of the structures) and third-most dominant (6% of the structures) clusters of their WT Aβ-42 simulation ensemble. An initial visualization (using SwissPDB) of our dominant WT structure indicated the presence of a β-hairpin in region ^{29}**GAII**GLM**VGGV**^{39}. This agreement suggests consistency between both studies. However, upon analysis of our dominant cluster's contact map (Fig. 4.4B, which contrasts with the pattern observed for an ordered β-hairpin; e.g., the C-terminal region of Fig. 4.4A), we only detected the presence of periodic β-bridges between RES. 29-31 and RES. 35-40. In Fig. 4.4B we used the VMD model of our dominant structure, which accurately captures the prevailing features of this cluster.

Regarding the GM6 monomer, the central structures and contact maps of the two major clusters (Figs. 4.3C and 4.4C) show a well preserved N-terminal β-hairpin spanning the ^{3}**EFR**HDS**GYE**^{11} segment, and a conserved CHC. Given the reduced motion displayed by the N-terminal region of this peptide as detected by the stability analysis of the previous section, we examined the remaining clusters to detect that presence of an N-terminal β-hairpin. Interestingly, 90% of the ensemble (clusters 1, 3, 4 and 7 in Fig. 4.3C) displayed this feature. This is therefore the most distinctive and predominant structural difference found between the soluble GM6 peptide and the two insoluble monomers (Dutch and WT). Additional features common to both clusters are: *i)* well preserved contacts between either E22 or D23 and K28; and *ii)* the extensive number of contacts between either E22 or K28 and RES. 1-20, which helps

hold the peptide's N-terminal half together and contributes to its reduced flexibility. Moreover, we found that 40% of the ensemble structures form a β-strand in RES. 27-29 that constantly interacts with the RES. 3-5 β-strand the region.

The two dominant clusters of the GM6 monomer also reveal some of their different structural characteristics. Concerning the C-terminal, while the dominant cluster establishes numerous short-lived contacts across its disordered U-shaped C-terminal, the second-most dominant group displays an extended C-terminal linked to a collapsed region comprising RES. 20-32.

IV. DISCUSSION

Understanding the structural dynamics of the Aβ monomers is important to aid the design of selective therapeutics that can prevent their oligomerization and the resulting toxicity of these aggregates. However, despite numerous experimental and computational efforts, there is still not a clear picture of the key structural features that seed the aggregation process. On the one hand, experimentally attaining high resolution structures has been hindered by the Aβ's fast aggregation rate; on the other hand, *in silico* modeling of the full peptide in explicit water still presents a significant computational challenge. Alternatively, given that the core structure of Aβ fibrils excludes the monomers' N-terminal,[65] many groups have focused on shorter/less amyloidogenic segments of the 11-40 (or 11-42) Aβ fragment. Remarkably, increasing evidence supporting a critical role of the N-terminal in Aβ aggregation has been reported by various experimental groups that have observed inhibition of oligomerization and fibril disaggregation upon N-terminal (mainly RES. 1-10) antibody or ligand binding.[66-71] Moreover, these studies concur that aggregation is not appreciably inhibited when anti-Aβ antibodies or ligands specific for the C-terminal or central Aβ region are used.

Using atomistic peptide models in explicit solvent, in this study we elucidated key structural differences between three Aβ-42 peptides, namely the wildtype, a soluble (GM6) and a highly insoluble (Dutch) variant. Specifically, the markers that we used for structural characterization indicate that the N-terminal (RES. 1-10) stability of these monomers correlates inversely with their relative aggregation tendency. This behavior contrasts with the one observed for the CHC and C-terminal regions, for which comparable structural dynamics are observed.

Our simulations show that the N-terminal region of the GM6 mutant forms a well conserved β-hairpin motif that significantly stabilizes this segment relative to that of the WT and Dutch peptides. Furthermore, despite being devoid of secondary structure, the N-terminal of the WT monomer still displays restricted motion when compared to that of the Dutch mutant, likely due to conserved interactions in the region encompassing RES. 3-18 (e.g., H-bonds between E3 and K16 or L17). These results are consistent with our N-terminal $\overline{\mathrm{RMSF}}_{\mathrm{back}}$ and $\mathrm{SASA}_{\mathrm{H}\phi}$ analyses. Our observations are also in line with those of a former study that used an analogous Aβ model,[36] regarding a decreased flexibility in the N-terminal region of the "more soluble" WT Aβ-40 variant, relative to that of WT Aβ-42. Those authors remark that the WT Aβ-40 monomer forms a small helical structure that stabilizes the N-terminal. Thus, the increased solubility of WT Aβ-40 over WT Aβ-42 and similarly, that of GM6 over WT Aβ-40, may be primarily due to an increase in N-terminal stability promoted by a better conserved motif in this region.

Unlike the N-terminal, the CHC and C-terminal regions of the three monomers show no clear distinguishing traits among them. All peptides display structured motifs at all or part of the CHC, with an occurrence of 58% (Dutch), 70% (WT) and 63% (GM6) in the ensemble of structures analyzed; these are predominantly 3_{10} helices (all variants) and occasionally β-sheets or β-bridges (WT and Dutch). This is consistent

with the peptides' comparable $RMSF_{back}$ values observed for this region. The CHC has been proposed as a site for aggregation initiation, due to potential destabilization of its helical structure and an increased exposure to solvent at pH > 6.[19] Our results show no appreciable changes in structure or $SASA_{H\phi}$ among the CHC of Aβ-42 monomers with widely varying solubilities. Nevertheless, this region has been previously identified as a major modulator of Aβ aggregation rate,[72] likely due to its repeated interactions with the remainder of the peptide (see Fig. 4.4). Moreover, the CHC is contained within the fibril forming core and may be responsible for the α−β transition observed during aggregation/fibrillization.[17] Thus, it is also likely that this region plays an important role in the Aβ's oligomerization pathways.

The C-terminal region of WT Aβ-42 has been found to possess less flexibility than that of WT Aβ-40,[16,63,73,74] a disparity that has led to the conjecture that Aβ aggregation may be seeded in the C-terminal and facilitated by the formation of stabilizing β-sheets in the 42-residue peptide,[35,36] which lower the entropic cost for aggregation. Our results indicate that the insoluble Aβ-42 variants, unlike the GM6 monomer, have propensity towards the formation of β-hairpin motifs (more so the Dutch mutant) in their C-terminal. However, we found no statistically significant difference in C-terminal $RMSF_{back}$ or $SASA_{H\phi}$ between the three cases, suggesting comparable flexibility and hydrophobicity between monomers of widely varying solubility. These later observations appear inconsistent with the behavior that would be expected for a simplistic C-terminal hydrophobic tail aggregation mechanism.

Overall, this work takes a step forward toward the identification of structural traits of Aβ monomers that can help clarify the effectiveness of anti-Aβ antibodies specific for this peptide's N-terminal (RES. 1-10) epitope[75] in reducing cerebral Aβ deposition in clinical trials. Our results suggest that aggregation is inhibited when the N-terminal region of Aβ is stabilized upon binding of N-terminal specific anti-Aβ

antibodies and ligands. The latter argument is consistent with conventional energetic premises suggesting that more *native-like* structures with increased stability and a lower exposed hydrophobic surface (e.g., GM6) would have reduced probability of association, and encounter higher energetic barriers when undergoing any conformational change taking place during oligomerization.

Conversely, the notion that Aβ N-terminal stabilization leads to reduced oligomerization does not necessarily imply that the N-terminal will be a seed for Aβ aggregation if this region's were unstable. In fact, experimental studies discourage this possibility, by reporting that: *i)* the N-terminal region is not part of the fibril forming core[65] and is still accessible for N-terminal specific anti-Aβ therapy within a fibrillar arrangement;[68] and *ii)* anti-Aβ antibodies also recognize Aβ's N-terminal binding region in oligomers,[67] suggesting an exposed N-terminal.

Despite being an improbable oligomerization site, the N-terminal likely acts as a "catalyst" of aggregation when it is unstable. Furthermore, we anticipate that the stabilization of the N-terminal leads to the formation of strong contacts between this region and RES.22-30 (e.g., D1-E22, E3-K28, F4-G29, D7-N27, H6-N27 in the GM6 variant), replacing weaker bonds that allow structural changes required for aggregation.

Increasing evidence suggests that the CHC, C and N terminals play an important role in what appears to be a stepwise transition with multiple oligomerization pathways.[3,6] Given that these key regions interplay in the aggregation process, an accurate identification of plausible oligomerization mechanisms entails the exploration of a very complex multidimensional conformational space; this is true even for the simplest case (i.e., dimerization). Initial approximations such as rigid backbone docking calculations are useful only if the interacting structures are representative of the Aβ ensemble of interest. Moreover, for flexible peptides such as

WT Aβ-42, binding may be optimized through conformational changes that make rigid body docking analyses inadequate. Thus, a sensible direction of future research on Aβ peptide dynamics likely involves an accurate determination of the structure (e.g., crystallization) of more soluble mutants, such as GM6, that can greatly assist in the validation/improvement of current *in silico* models, which in turn can be used to improve the structural prediction of the WT Aβ monomers used for further oligomerization analyses.

Ongoing efforts on this topic are focused on the study of the dynamics of homo/heterodimers assembled from representative structures of the three variants analyzed in this work.

ACKNOWLEDGEMENTS

The authors are grateful for support from the National Science Foundation, Award No. CBET 0933092, and acknowledge Corning Inc. for kindly providing computational resources that allowed the completion of this work.

REFERENCES

1. Haass, C. & Selkoe, D.J. Soluble protein oligomers in neurodegeneration: lessons from the Alzheimer's amyloid β-peptide. *Nat Rev Mol Cell Biol* **8**, 101 (2007).

2. Shankar, G.M. et al. Amyloid-β protein dimers isolated directly from Alzheimer's brains impair synaptic plasticity and memory. *Nat Med* **14**, 837 (2008).

3. Bitan, G. Amyloid beta-protein (Abeta) assembly: Abeta 40 and Abeta 42 oligomerize through distinct pathways. *Proceedings of the National Academy of Sciences* **100**, 330 (2002).

4. Walsh, D.M., Tseng, B.P., Rydel, R.E., Podlisny, M.B. & Selkoe, D.J. The Oligomerization of Amyloid β-Protein Begins Intracellularly in Cells Derived from Human Brain†. *Biochemistry* **39**, 10831 (2000).

5. Kayed, R. et al. Common Structure of Soluble Amyloid Oligomers Implies Common Mechanism of Pathogenesis. *Science* **300**, 486 (2003).

6. Bernstein, S.L. et al. Amyloid-β protein oligomerization and the importance of tetramers and dodecamers in the aetiology of Alzheimer's disease. *Nature Chem* **1**, 326 (2009).

7. Ward, R.V. et al. Fractionation and characterization of oligomeric, protofibrillar and fibrillar forms of beta-amyloid peptide. *Biochem. J.* **348**, 137 (2000).

8. Zhang, S. The Alzheimer's Peptide Aβ Adopts a Collapsed Coil Structure in Water. *Journal of Structural Biology* **130**, 130 (2000).

9. Hou, L. et al. Solution NMR Studies of the Aβ(1−40) and Aβ(1−42) Peptides Establish that the Met35 Oxidation State Affects the Mechanism of Amyloid Formation. *J. Am. Chem. Soc.* **126**, 1992 (2004).

10. Kirkitadze, M.D., Condron, M.M. & Teplow, D.B. Identification and characterization of key kinetic intermediates in amyloid β-protein fibrillogenesis1. *Journal of Molecular Biology* **312**, 1103 (2001).

11. Lazo, N.D., Grant, M.A., Condron, M.C., Rigby, A.C. & Teplow, D.B. On the nucleation of amyloid β-protein monomer folding. *Protein Sci.* **14**, 1581 (2005).

12. Wood, S.J., Wetzel, R., Martin, J.D. & Hurle, M.R. Prolines and Aamyloidogenicity in Fragments of the Alzheimer's Peptide .beta./A4. *Biochemistry* **34**, 724 (1995).

13. Szabo, Z. et al. An FT-IR Study of the [beta]-Amyloid Conformation: Standardization of Aggregation Grade. *Biochemical and Biophysical Research Communications* **265**, 297 (1999).

14. Crescenzi, O. et al. Solution structure of the Alzheimer amyloid beta-peptide (1-42) in an apolar microenvironment. Similarity with a virus fusion domain. *Eur J Biochem* **269**, 5642 (2002).

15. Sticht, H. et al. Structure of amyloid A4-(1--40)-peptide of Alzheimer's disease. *European Journal of Biochemistry* **233**, 293 (1995).

16. Lim, K.H., Henderson, G.L., Jha, A. & Louhivuori, M. Structural, Dynamic Properties of Key Residues in Aβ Amyloidogenesis: Implications of an Important Role of Nanosecond Timescale Dynamics. *ChemBioChem* **8**, 1251 (2007).

17. Tomaselli, S. et al. The α-to-β Conformational Transition of Alzheimer's Aβ-(1–42) Peptide in Aqueous Media is Reversible: A Step by Step Conformational Analysis Suggests the Location of β Conformation Seeding. *ChemBioChem* **7**, 257 (2006).

18. Chebaro, Y., Mousseau, N. & Derreumaux, P. Structures and Thermodynamics of Alzheimer's Amyloid-β Aβ(16−35) Monomer and Dimer by Replica Exchange Molecular Dynamics Simulations: Implication for Full-Length Aβ Fibrillation. *J. Phys. Chem. B* **113**, 7668 (2009).

19. Khandogin, J. & Brooks, C.L. Linking Folding with Aggregation in Alzheimer's β-Amyloid Peptides. *Proceedings of the National Academy of Sciences of*

the United States of America **104**, 16880 (2007).

20. Cecchini, M., Curcio, R., Pappalardo, M., Melki, R. & Caflisch, A. A Molecular Dynamics Approach to the Structural Characterization of Amyloid Aggregation. *Journal of Molecular Biology* **357**, 1306 (2006).

21. Jang, S. & Shin, S. Computational Study on the Structural Diversity of Amyloid Beta Peptide (Aβ10-35) Oligomers. *The Journal of Physical Chemistry B* **112**, 3479 (2008).

22. Anand, P., Nandel, F.S. & Hansmann, U.H.E. The Alzheimer β-amyloid (Aβ[sub 1–39]) dimer in an implicit solvent. *J. Chem. Phys.* **129**, 195102 (2008).

23. Cecchini, M., Rao, F., Seeber, M. & Caflisch, A. Replica exchange molecular dynamics simulations of amyloid peptide aggregation. *J. Chem. Phys.* **121**, 10748 (2004).

24. Baumketner, A., Krone, M.G. & Shea, J. Role of the familial Dutch mutation E22Q in the folding and aggregation of the 15–28 fragment of the Alzheimer amyloid-β protein. *Proceedings of the National Academy of Sciences* **105**, 6027 (2008).

25. Wei, G. & Shea, J. Effects of Solvent on the Structure of the Alzheimer Amyloid-β(25–35) Peptide. *Biophysical Journal* **91**, 1638 (2006).

26. Baumketner, A. et al. Structure of the 21-30 fragment of amyloid β-protein. *Protein Sci.* **15**, 1239 (2006).

27. Baumketner, A. & Shea, J. The Structure of the Alzheimer Amyloid β 10-35 Peptide Probed through Replica-Exchange Molecular Dynamics Simulations in Explicit Solvent. *Journal of Molecular Biology* **366**, 275 (2007).

28. Daidone, I. et al. β-Hairpin conformation of fibrillogenic peptides: Structure and α-β transition mechanism revealed by molecular dynamics simulations. *Proteins* **57**, 198 (2004).

29. Krone, M.G. et al. Effects of Familial Alzheimer's Disease Mutations on the

Folding Nucleation of the Amyloid β-Protein. *Journal of Molecular Biology* **381**, 221 (2008).

30. Nguyen, P.H., Li, M.S., Stock, G., Straub, J.E. & Thirumalai, D. Monomer adds to preformed structured oligomers of Abeta-peptides by a two-stage dock-lock mechanism. *Proceedings of the National Academy of Sciences* **104**, 111 (2007).

31. Huet, A. & Derreumaux, P. Impact of the Mutation A21G (Flemish Variant) on Alzheimer's β-Amyloid Dimers by Molecular Dynamics Simulations. *Biophysical Journal* **91**, 3829 (2006).

32. Han, W. & Wu, Y. Molecular dynamics studies of hexamers of amyloid-β peptide (16–35) and its mutants: Influence of charge states on amyloid formation. *Proteins* **66**, 575 (2007).

33. Kassler, K., Horn, A.H.C. & Sticht, H. Effect of pathogenic mutations on the structure and dynamics of Alzheimer's Aβ42-amyloid oligomers. *J Mol Model* **16**, 1011 (2009).

34. Baumketner, A. et al. Amyloid beta-protein monomer structure: A computational and experimental study. *Protein Science* **15**, 420 (2006).

35. Yang, M. & Teplow, D. Amyloid β-Protein Monomer Folding: Free-Energy Surfaces Reveal Alloform-Specific Differences. *Journal of Molecular Biology* **384**, 450 (2008).

36. Sgourakis, N.G., Yan, Y., McCallum, S.A., Wang, C. & Garcia, A.E. The Alzheimer's Peptides Aβ40 and 42 Adopt Distinct Conformations in Water: A Combined MD / NMR Study. *Journal of Molecular Biology* **368**, 1448 (2007).

37. Luttmann, E. & Fels, G. All-atom molecular dynamics studies of the full-length β-amyloid peptides. *Chemical Physics* **323**, 138 (2006).

38. Flöck, D., Colacino, S., Colombo, G. & Di Nola, A. Misfolding of the amyloid β-protein: A molecular dynamics study. *Proteins* **62**, 183 (2006).

39. Triguero, L., Singh, R. & Prabhakar, R. Molecular Dynamics Study To Investigate the Effect of Chemical Substitutions of Methionine 35 on the Secondary Structure of the Amyloid β (Aβ(1−42)) Monomer in Aqueous Solution. *J. Phys. Chem. B* **112**, 2159 (2008).

40. Kaminski, G.A., Friesner, R.A., Tirado-Rives, J. & Jorgensen, W.L. Evaluation and Reparametrization of the OPLS-AA Force Field for Proteins via Comparison with Accurate Quantum Chemical Calculations on Peptides†. *The Journal of Physical Chemistry B* **105**, 6474 (2001).

41. Jorgensen, W.L., Chandrasekhar, J., Madura, J.D., Impey, R.W. & Klein, M.L. Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics* **79**, 926 (1983).

42. Petkova, A.T. A structural model for Alzheimer's beta -amyloid fibrils based on experimental constraints from solid state NMR. *Proceedings of the National Academy of Sciences* **99**, 16742 (2002).

43. Sugita, Y. & Okamoto, Y. Replica-exchange molecular dynamics method for protein folding. *Chemical Physics Letters* **314**, 141 (1999).

44. Brenner, P., Sweet, C.R., VonHandorf, D. & Izaguirre, J.A. Accelerating the replica exchange method through an efficient all-pairs exchange. *Journal of Chemical Physics* **126**, 074103 (2007).

45. Wurth, C., Guimard, N.K. & Hecht, M.H. Mutations that Reduce Aggregation of the Alzheimer's A[beta]42 Peptide: an Unbiased Search for the Sequence Determinants of A[beta] Amyloidogenesis. *Journal of Molecular Biology* **319**, 1279 (2002).

46. Fisher, A.C., Kim, W. & Delisa, M.P. Genetic selection for protein solubility enabled by the folding quality control feature of the twin-arginine translocation pathway. *Protein Science* **15**, 449 (2006).

47. Kim, W. & Hecht, M.H. Generic hydrophobic residues are sufficient to promote aggregation of the Alzheimer's Abeta42 peptide. *Proceedings of the National Academy of Sciences* **103**, 15824 (2006).

48. Murakami, K. et al. Neurotoxicity and Physicochemical Properties of Aβ Mutant Peptides from Cerebral Amyloid Angiopathy. *Journal of Biological Chemistry* **278**, 46179 (2003).

49. Pérez, A., Morelli, L., Cresto, J.C. & Castaño, E.M. Degradation of soluble amyloid beta-peptides 1-40, 1-42, and the Dutch variant 1-40Q by insulin degrading enzyme from Alzheimer disease and control brains. *Neurochem. Res* **25**, 247 (2000).

50. Davis, J. & Van Nostrand, W.E. Enhanced pathologic properties of Dutch-type mutant amyloid beta-protein. *Proceedings of the National Academy of Sciences of the United States of America* **93**, 2996 (1996).

51. Fraser, P.E. et al. Fibril formation by primate, rodent, and Dutch-hemorrhagic analogs of Alzheimer amyloid .beta.-protein. *Biochemistry* **31**, 10716-10723 (1992).

52. Waal, R.M.W.D., Schipper, J.J. & Nostrand, W.E.V. Rapid Degeneration of Cultured Human Brain Pericytes by Amyloid β-Protein. *Journal of Neurochemistry* **68**, 1135 (1997).

53. Irie, K. et al. Structure of [beta]-amyloid fibrils and its relevance to their neurotoxicity: Implications for the pathogenesis of Alzheimer's disease. *Journal of Bioscience and Bioengineering* **99**, 437 (2005).

54. Massi, F. & Straub, J.E. Probing the Origins of Increased Activity of the E22Q "Dutch" Mutant Alzheimer's [beta]-Amyloid Peptide. *Biophysical Journal* **81**, 697 (2001).

55. Cruz, L. et al. Solvent and mutation effects on the nucleation of amyloid β-protein folding. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 18258 (2005).

56. Spoel, D.V.D. et al. GROMACS: Fast, flexible, and free. *Journal of Computational Chemistry* **26**, 1701 (2005).

57. Velez-Vega, C., Fenwick, M.K. & Escobedo, F.A. Simulated Mutagenesis of the Hypervariable Loops of a Llama VHH Domain for the Recovery of Canonical Conformations. *J. Phys. Chem. B* **113**, 1785 (2009).

58. Nosé, S. A unified formulation of the constant temperature molecular dynamics methods. *J. Chem. Phys.* **81**, 511 (1984).

59. Hoover, W. Canonical dynamics: Equilibrium phase-space distributions. *Phys. Rev. A* **31**, 1695 (1985).

60. Hess, B., Bekker, H., Berendsen, H.J.C. & Fraaije, J.G.E.M. LINCS: A linear constraint solver for molecular simulations. *Journal of Computational Chemistry* **18**, 1463 (1997).

61. Humphrey, W., Dalke, A. & Schulten, K. VMD: Visual molecular dynamics. *Journal of Molecular Graphics* **14**, 33 (1996).

62. Guex, N. & Peitsch, M.C. SWISS-MODEL and the Swiss-Pdb Viewer: An environment for comparative protein modeling. *Electrophoresis* **18**, 2714 (1997).

63. Riek, R., Guntert, P., Dobeli, H., Wipf, B. & Wuthrich, K. NMR studies in aqueous solution fail to identify significant conformational differences between the monomeric forms of two Alzheimer peptides with widely different plaque-competence Aβ(1-40)^{ox} and Aβ(1-42)^{ox}. *European Journal of Biochemistry* **268**, 5930 (2001).

64. Dobson, C.M. Protein misfolding, evolution and disease. *Trends in Biochemical Sciences* **24**, 329 (1999).

65. Luhrs, T. 3D structure of Alzheimer's amyloid- (1-42) fibrils. *Proceedings of the National Academy of Sciences* **102**, 17342 (2005).

66. Orner, B.P., Liu, L., Murphy, R.M. & Kiessling, L.L. Phage Display Affords Peptides that Modulate β-Amyloid Aggregation. *J. Am. Chem. Soc.* **128**, 11882

(2006).

67. Frenkel, D., Balass, M. & Solomon, B. N-terminal EFRH sequence of Alzheimer's [beta]-amyloid peptide represents the epitope of its anti-aggregating antibodies. *Journal of Neuroimmunology* **88**, 85 (1998).

68. Solomon, B., Koppel, R., Frankel, D. & Hanan-Aharon, E. Disaggregation of Alzheimer β-amyloid by site-directed mAb. *Proceedings of the National Academy of Sciences of the United States of America* **94**, 4109 (1997).

69. Solomon, B., Koppel, R., Hanan, E. & Katzav, T. Monoclonal antibodies inhibit in vitro fibrillar aggregation of the Alzheimer beta-amyloid peptide. *Proceedings of the National Academy of Sciences of the United States of America* **93**, 452 (1996).

70. Bard, F. et al. Epitope and isotype specificities of antibodies to β-amyloid peptide for protection against Alzheimer's disease-like neuropathology. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 2023 (2003).

71. Lee, M. et al. Abeta42 immunization in Alzheimer's disease generates Abeta N-terminal antibodies. *Annals of Neurology* **58**, 430 (2005).

72. de Groot, N.S., Aviles, F.X., Vendrell, J. & Ventura, S. Mutagenesis of the central hydrophobic cluster in Abeta42 Alzheimer's peptide. Side-chain properties correlate with aggregation propensities. *FEBS Journal* **273**, 658 (2006).

73. Yan, Y. & Wang, C. Aβ42 is More Rigid than Aβ40 at the C Terminus: Implications for Aβ Aggregation and Toxicity. *Journal of Molecular Biology* **364**, 853 (2006).

74. Yan, Y., Liu, J., Mccallum, S., Yang, D. & Wang, C. Methyl dynamics of the amyloid-β peptides Aβ40 and Aβ42. *Biochemical and Biophysical Research Communications* **362**, 410 (2007).

75. Szabo, P., Relkin, N. & Weksler, M. Natural human antibodies to amyloid beta peptide. *Autoimmunity Reviews* **7**, 415 (2008).

APPENDIX A

For validation of the sampling method and force field implemented in this study, several supportive runs were performed.

Three MD runs at 300°K were carried out with mutant 3-FFSa (see Table 1.1 for the mutations made in each case), starting from non-canonical H1 conformations with RMSD values of 1.56-1.66 Å from 1DFB, and side chains F27 and F29 pointing outwards toward the solvent. These MD runs used either CHARMM22-CMAP with GBSW implicit solvent, CHARMM19 with GBMVA implicit solvent, or CHARMM22-CMAP with 7662 molecules of TIP3 water. None of the MD simulations were able to achieve the expected burial of the H1 phenylalanine side chains, nor achieve H1 RMSD values below 1.2 Å from 1DFB, when evaluated for a period of 20 ns. These results suggest a need for improved conformational sampling, which was achieved by the use of REM.

Additional 5-ns validation simulations were carried out for the wildtype, 1-Fa, 2-FF and 3-FFSa mutants, using REM (spanning a range of 300-900 K), with CHARMM22-CMAP and CHARMM19 force fields. GBMVA and EEF were explored for CHARMM19, and GBSW was implemented with CHARMM22-CMAP. It was found that none of the runs using either CHARMM22-CMAP/GBSW or CHARMM19/EEF1 achieved the timely hydrophobic burial for this system. Conversely, the simulations with CHARMM19/GBMVA achieved conformations with buried F27 and F29 residues early in the runs. Despite being an older force field, CHARMM19 may be more appropriate for certain systems such as the one studied here.

Despite being unable to achieve burial of loop side chains when using CHARMM22-CMAP/GBSW, the possibility was explored that this scheme could verify the stability of the canonical H1 loop conformations obtained from the

mutational analysis presented. Three 5-ns REM (300-900°K) runs were performed under the mentioned scheme, starting from 3-FFSa, 2-FF and 1-Fa structures with low (0.8-1 Å) H1 RMSD values from 1DFB and H1 hydrophobic side chains buried within the loop. For the 3-FFSa case, it was found that the H1 and H2 loops have low variability, remaining close to the initial conformation and with the H1 loop F27 & F29 residues buried in the proper positions. 85% and 100% of the H1 and H2 loop conformations respectively have RMSD values below 1.2 Å from 1DFB and 1FVC, supporting the high stability of the 3-FFSa structure obtained using CHARMM19/GBMVA. On the other hand for the 1-Fa and 2-FF cases, 0% / 48% (H1) and 95% / 100% (H2) of the configurations present RMSD values below 1.2 Å from 1DFB and 1FVC, respectively. Both cases show a loop stability comparable to that found in the REM simulations presented in the results section.

As an additional control experiment for the REM setup used in the mutational analysis presented, a 64 amino acid reduced model of camel antibody cAb-CA05 (PDB code: 1F2X) was simulated for 5 ns, using the same REM setup and force field/solvation model implemented for the 1HCV wildtype and mutant runs discussed in the paper. This VHH is particularly appealing for validation, given that its crystal structure shows an H1 type 1 canonical conformation and a non-canonical H2 conformation. This structure has a 'Y' and 'V' at positions 27 and 29, respectively. Since the hydropathy index of 'V' is higher than that of 'F', we might expect the H1 loop of this antibody to be highly stable. For the complete run, it is observed that the H1 loop closely maintains its initial type 1 structure, with 93% of the configurations having a RMSD value below 1.2 Å. It is noted that various replicas that occupied the high temperatures with H1 loop RMSD values of 3-4 Å were gradually able to diffuse to the temperature of interest where they displayed the expected type 1 H1 canonical. With respect to the H2 region, the results agree well with the non-canonical nature of

this loop (for the simulation time considered), with 63% of its conformations having a RMSD value below 1.2 Å from the minimized (initial) structure (which does not match any of the canonical types that have been previously identified).

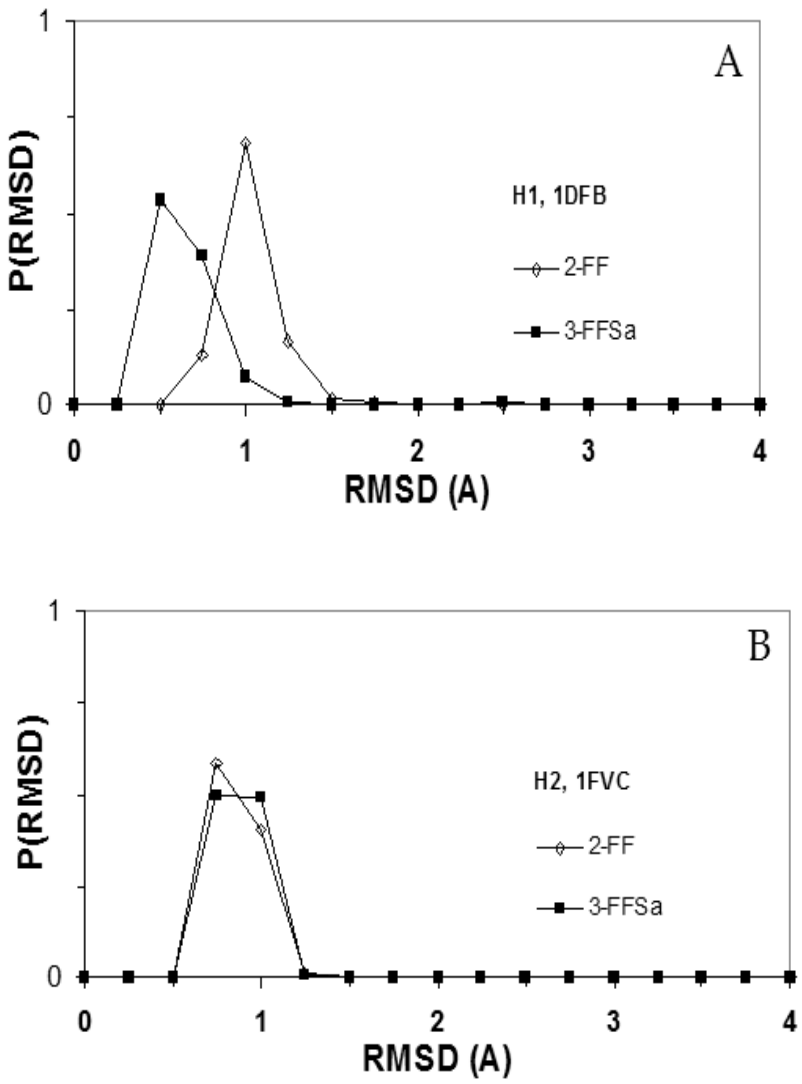


Figure. A.1. Probabilities of occurrence for the RMSD values of the simulated mutants 2-FF and 3-FFSa for H1 (*A*) and H2 (*B*) loops at 300°K relative to reference structures 1DFB and 1FVC. The distributions account for the data evaluation period of a 10-ns run.

APPENDIX B

Consider a certain mutant for which conventional MD REM[46] is used. Such system is studied within a generalized ensemble formed by M replicas of the same mutant, each simulated at a different temperature T. Since the replicas are non-interacting, the weight factor W_{REM} for state X is given by the product of the Boltzmann factors for each replica.

$$W_{REM}(X) = \exp\left\{-\sum_{i=1}^{M} \beta_{m(i)} H\left(q^{[i]}, p^{[i]}\right)\right\} \tag{25}$$

In Eq. (25), H is the Hamiltonian, q and p are respectively the set of coordinates and momenta for the atoms in replica i, and β_m is the inverse temperature of replica i, which has a one-to-one correspondence with temperature m. Periodic exchanges between replicas i and j at T_m and T_n, respectively, i.e.,

$$X = \left(\ldots, x_m^{[i]}, \ldots, x_n^{[j]}, \ldots\right) \rightarrow X' = \left(\ldots, x_m^{[j]'}, \ldots, x_n^{[i]'}, \ldots\right) \tag{26}$$

are performed to facilitate convergence towards an equilibrium distribution. For this purpose, the detailed balance condition is imposed on the transition probability $w(X \rightarrow X')$,

$$W_{REM}(X) w(X \rightarrow X') = W_{REM}(X') w(X' \rightarrow X) \tag{27}$$

The latter condition is satisfied by the Metropolis acceptance criterion,

$$P_{acc}\left(X\left(\ldots, x_m^{[i]}, \ldots, x_n^{[j]}, \ldots\right) \rightarrow X'\left(\ldots, x_m^{[j]'}, \ldots, x_n^{[i]'}, \ldots\right)\right) = \begin{cases} 1, & for \quad \Delta \le 0 \\ \exp(-\Delta), & for \quad \Delta > 0 \end{cases} \tag{28}$$

$$\text{where} \quad \Delta = \left(\beta_m - \beta_n\right) \times \left(U_j - U_i\right) \tag{29}$$

In Eq. (29), U_i and U_j are the potential energies of configurations *i* and *j*, with corresponding inverse temperatures β_m and β_n. The kinetic energy terms have been eliminated through velocity rescaling.

Consider now an exchange event between temperatures *m* and *n* of respective mutants A and B, using a criterion analogous to Eq. (28). For this multiple-mutant replica exchange method (MMREM), detailed balance condition on the transition probability leads to:

$$\frac{w(X \to X')}{w(X' \to X)} = \frac{w_A(q_A \mid \beta_n) w_B(q_B \mid \beta_m)}{w_A(q_A \mid \beta_m) w_B(q_B \mid \beta_n)} \tag{30}$$

where the probability weight for any given mutant is given by,

$$w(q \mid \beta) = \exp(-\beta U(q)) / Q(\beta, H) \tag{31}$$

In Eq. (31) *Q*, *H* and *U* are the partition function, Hamiltonian and energy of the system, respectively. Substituting the appropriate expressions of Eq. (31) into Eq. (30) and rearranging we have,

$$\frac{w(X \to X')}{w(X' \to X)} = \frac{\exp\left[-(\beta_m - \beta_n)(U_B(q_B) - U_A(q_A))\right]}{\frac{Q(\beta_n, H_A)}{Q(\beta_m, H_A)} \frac{Q(\beta_m, H_B)}{Q(\beta_n, H_B)}} \tag{32}$$

Now, given that mutant A(B) (read A or B) is simulated in the canonical ensemble, its partition functions at T_m and T_n are related to the Helmholtz free energy **A** by the expressions:

$$\begin{aligned} \beta_m \mathrm{A}_m^{A(B)} &= -\ln Q(\beta_m, H_{A(B)}) \\ \beta_n \mathrm{A}_n^{A(B)} &= -\ln Q(\beta_n, H_{A(B)}) \end{aligned} \tag{33}$$

We can define a change in free energy $\Delta F^{A(B)}$ associated with a virtual temperature swap for mutant A(B) by subtracting the second equality of Eq. (33) from the first one:

$$\Delta F^{A(B)} = F_n^{A(B)} - F_m^{A(B)} = \beta_n \mathbf{A}_n^{A(B)} - \beta_m \mathbf{A}_m^{A(B)} = -\ln \frac{Q(\beta_n, H_{A(B)})}{Q(\beta_m, H_{A(B)})} \tag{34}$$

Equation (34) can then be introduced into equation 8 for both mutants A and B, and the result simplified to:

$$\frac{w(X \to X')}{w(X' \to X)} = \exp\left[-(\beta_m - \beta_n)(U_B(q_B) - U_A(q_A)) + (\Delta F^A - \Delta F^B)\right] \tag{35}$$

Equation (35) may be satisfied by the Metropolis acceptance criterion (equation 4), but with Δ now defined by:

$$\Delta_{MMREM} = (\beta_m - \beta_n)(U_B(q_B) - U_A(q_A)) - (\Delta F^A - \Delta F^B) \tag{36}$$

If mutants A and B were identical, then $\Delta F^A - \Delta F^B = 0$ and Δ_{MMREM} would reduce to the Δ of the conventional REM [see Eq. (29)]. The free energy change (ΔF) associated with the temperature change for a given mutant can be evaluated by using the acceptance ratio method originally proposed by Bennett;[67] e.g., applying the "unoptimized" version of the method to our system leads to:

$$\Delta F^{A(B)} = -\ln\left[\frac{< P_{acc}^{A(B)}(\beta_m \to \beta_n) >_{\beta_m}}{< P_{acc}^{A(B)}(\beta_n \to \beta_m) >_{\beta_n}}\right] \tag{37}$$

In Eq. (37), the $\langle\rangle$ brackets denote ensemble averages taken at the subscripted β, and the acceptance probability (for the virtual temperature changes) can be found, for

example, by application of Barker's rule [68] to obtain $P_{acc}^{A(B)} = e^{-(\beta_n-\beta_m)U_{A(B)}} \Big/ \left(1+e^{-(\beta_n-\beta_m)U_{A(B)}}\right)$. These $P_{acc}^{A(B)}$ values can be readily obtained (at no cost) from the data of the exchange attempts in the MMREM run. The $\Delta F^{A(B)}$ values would be initialized to zero at the beginning of the MMREM run, calculated on the fly, but updated only after a minimum of statistics have been accumulated. Note that, besides requiring no computational overhead, evaluation of $\Delta F^{A(B)}$ can be of interest independent of MMREM, e.g., for thermodynamic analysis.

Rigorous execution of MMREM requires then the evaluation of the mutants' free energy changes [Eq. (37)] for subsequent calculation of Δ_{MMREM}. This procedure, however, may be inconvenient to implement when using some of the available computational packages. An alternative, albeit approximate approach for analyses such as the one implemented in this study, exploits the fact that point mutations introduce relatively small perturbations to the basal energy of the system. If mutants A and B differ only by a small number of mutations, it is plausible to assume that ΔF^{A}-ΔF^{B} in Eq. (36) will be small relative to the first term. Under these conditions, we can approximate $\Delta_{\text{MMREM}} \approx \Delta$ [with Δ as defined in Eq. (29)]; this variant was referred to as the "simplified" version in the main text. A quasi-rigorous version of the approach would entail separating the contributions to $U_{A(B)}$ into "common" interactions (among shared residues) and "mutant-specific" interactions and keeping constant the "temperature" of the latter; in such a case, ΔF^{A} would also approach ΔF^{B} as they would entail temperature changes of the common interactions only.

Ongoing research aims at evaluating the applicability of both MMREM schemes (rigorous and simplified versions) in varying scenarios.

APPENDIX C

COMPARATIVE ANALYSIS OF BG PROTOCOLS

The CBG method discussed in Sec. II C was introduced to prevent the exponential growth of trial runs as the system moves closer to state B. Here we compare the performance of CBG with those of other BG schemes (see below) for the estimation of the average transition rate constant. We selected as test-bed a particle moving on a two-dimensional potential energy surface which has been previously used to test path sampling methods.[33-34] Figure C.1 shows a contour graph of this energy landscape and the isocommittor surfaces used to partition the phase space (solid red lines). The reaction coordinate ($\lambda = p_B$) for this system was previously estimated in Ref. 33. Figure C.1 shows the two stable states defined by circles of radius 1.0 and centered at the minima: state A (-4,0) and state B (4,0). The kinetics of the system was simulated using Brownian dynamics at $\beta = 1/k_B T = 2.5$, particle mass *m*=1.0, friction coefficient $\gamma = 2.5$, and time increments $\Delta t = 0.1$. Reflective boundaries were used to keep the particle inside of the phase space region -8<*x*<8 and -4<*y*<8. Further details on the energy potential are given elsewhere.[33-34]

We studied three different branch growth schemes: (i) the original (BG) framework where the number of trial runs per state at λ_i is fixed to $k_i = 10$ (i.e., $k_i^j = k_i^{\max} = 10$), (ii) the constrained branch growth (CBG) protocol described in Sec. IIC in which k_i^j is given by Eq. (18) with $k_{\min} = 3$ and $M_i = 1000$, and (iii) the RBG scheme where the number of trial runs for each state *j* at λ_i is selected randomly from a number between $k_{\min} = 3$ and $k_i^{\max} = 10$ (i.e., $k_i^j = rand[k_{\min}, k_i^{\max}]$).

Because in the original BG scheme $k_i^j = k_i$ is constant at each interface λ_i , this protocol "automatically" harvests the correctly weighted TPE. Hence, paths for which more trial runs are successful produce more branches, making a larger contribution to the TPE.[16] However, when a variable approach (i.e., $k_i^{(j,m)} \neq k_i$

depends on the m^{th} run) is used to generate the TPE (e.g., the CBG and RBG schemes), the weight for each transition pathway must be multiplied by a factor W:

$$W^{(j,m)} = \prod_{i=1}^{n-1} \frac{k_i^{\max}}{k_i^{(j,m)}} \tag{38}$$

m is an index that denotes the run (i.e., starting from a randomly selected point at λ_0) such that $k_i^{(j,m)}$ is the number of trial runs started at interface λ_i for the successful trajectory j of run m. Hence, for a given run m, one can estimate the probability to reach B from λ_0 as:

$$P_m(\lambda_n \mid \lambda_0) = \left[\sum_{j=1}^{N_s^{(n-1,m)}} W^{(j,m)} \right] \times \frac{1}{\prod_{i=1}^{n-1} k_i^{\max}} = \sum_{j=1}^{N_s^{(n-1,m)}} \frac{1}{\prod_{i=1}^{n-1} k_i^{(j,m)}} \tag{39}$$

where $N_s^{(n-1,m)}$ is the number of successful trajectories (that reached λ_n) for the m^{th} run. Note that Eq. (39) simplifies to $P_m(\lambda_n \mid \lambda_0) = N_s^{n-1} / \prod_{i=1}^{n-1} k_i^{\max}$ for the original BG scheme.[16] Finally, the average of the $P(\lambda_n \mid \lambda_0)$ probability is estimated from:

$$P(\lambda_n \mid \lambda_0) = \left\langle P_m(\lambda_n \mid \lambda_0) \right\rangle = \frac{1}{N_0} \sum_{m=1}^{N_0} P_m(\lambda_n \mid \lambda_0) \tag{40}$$

where N_0 is the number of BG runs (points started at λ_0). Care should be taken when selecting the range for $k_i^{(j,m)}$ values, since Eq. (39) implies that paths generated by firing $k_i^{(j,m)} << k_i^{\max}$ have greater weight $W^{(j,m)}$, which can introduce statistical inefficiency in the $P_m(\lambda_n \mid \lambda_0)$ estimate. Moreover, Eq. (40) assigns an equal weight to all $P_m(\lambda_n \mid \lambda_0)$ estimates from Eq. (39) which may compromise accuracy if the k's differ considerably among runs. It is also expected that the precision of the p_B estimations will be affected when the k's are far from the $k_i^{(j,m)} \approx k_i$ condition, since

they are based on p_B estimates per point. Hence, when using the FFS-LSE method, p_B estimates for states along successful pathways should be weighted in accordance with Eq. (38).

The rate constant calculations were carried out as a series of 10 blocks, each one consisting of $N_0 = 100$ BG runs, where $P_m(\lambda_n \mid \lambda_0)$ for each successful transition from state A to B was estimated using Eq. (39) for the BG and RBG and Eq. (2) for the CBG. The $P(\lambda_n \mid \lambda_0)$ value for each block was estimated via Eq. (40) and the final $P(\lambda_n \mid \lambda_0)$ estimate was obtained by averaging over the 10 blocks. The results were compared to rate constants which were obtained by "brute force" using the mean passage time from 10 blocks with 1000 transitions per block.

We first studied the performance of the various BG schemes for a sub-optimal choice of the λ order parameter. For this purpose, we used the x-coordinate and the λ space was partitioned into n=9 interfaces positioned at λ_i ($0 \le i < n$-1): $\lambda(x)$ = {-3,-2.5,-2,-1.5,-1,-0.5,0,1,2}, and λ_A and $\lambda_{B=n}$ were taken as the circular regions enclosing states A and B (see Fig. C.1). In Figure C.2 (left), the rate constant values for the three protocols are compared using the brute-force value as a basis. It is seen that the original BG and the CBG schemes agree well with the brute force result, whereas the RBG scheme overestimates it. We then studied the case where λ is closer to the true reaction coordinate p_B. For this case, the transition region was partitioned with eight interfaces (n=8) positioned at λ_i ($0 \le i \le n$-1): p_B = {0.0,0.05,0.10,0.20,0.30,0.40,0.50,0.60}. Figure C.2 (right) shows that all three variations of the BG scheme provide good estimates of the rate constant. In this case, the statistical inefficiency caused by generating paths with different weight in the RBG scheme is reduced, because all states along any given isocommittor surface have similar probability to reach the next interface (i.e., ergodic sampling of the path space). In the original BG protocol $k_i^{(j,m)} = k_i$ for all the m runs, and so equal-weight

$P_m(\lambda_n \mid \lambda_0)$ estimates are obtained (i.e, all paths in the TPE have the same weight). For the CBG scheme, $k_i^{(j,m)} = k_i^m$ in the majority of the instances so that all the paths in the m^{th} run are expected to have a commensurate weight. In contrast, for the RBG scheme a suboptimal order parameter may lead to unduly large weights for rare trajectories having a small $k_i^{(j,m)}$ along the path.

Our first set of FFS runs for the Trp-cage N-L transition (described in Sections III and IV) used $RMSD_{hx}$ as order parameter (a good choice for λ) and the RBG protocol described above, with $k_0 = 10$, $k_{\min} = 3$, and $k_i^{\max} = 10$ at each interface. The RBG was used because at that time we had not yet developed the more effective CBG method. The RGB results served the purpose to produce a fast but preliminary exploration of the TPE, preparing the way for the second set of CBG simulations.

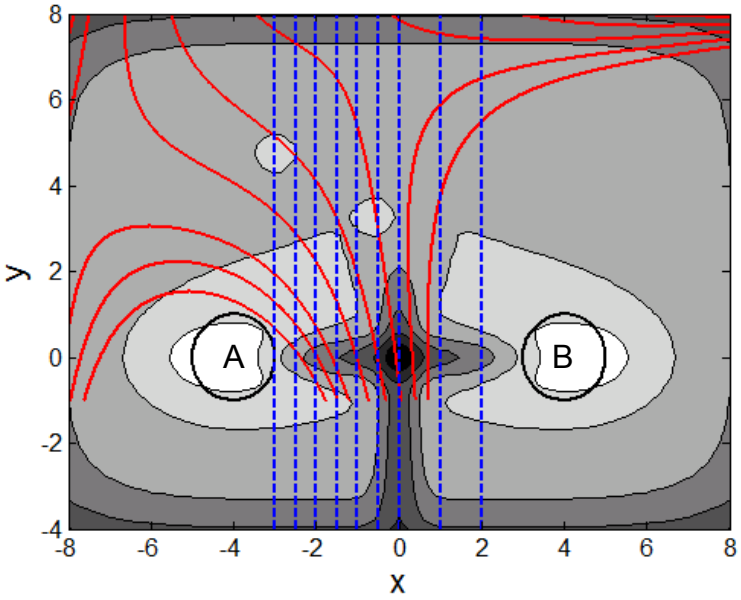


Figure C.1. Contour graph of the free energy surface for the two-dimensional potential.[33-34] The color scheme changes from highest (gray) to lowest (white) elevations. The basins are shown by the cycles labeled A, B. The order parameter and staging is also shown for the predicted p_B committors (solid red lines) from the reaction coordinate model found by FFS-LSE simulations[33], and x-coordinate (dotted blue lines).

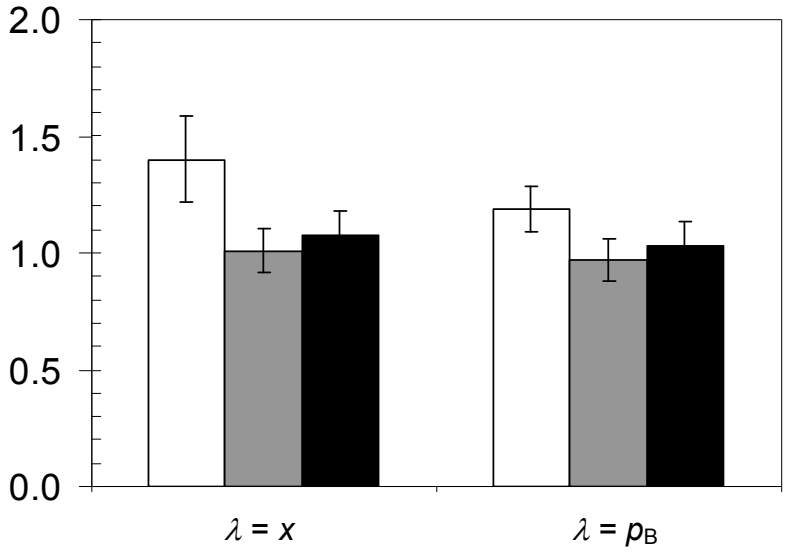


Figure C.2. Ratios between the rate constant found for different BG schemes and the one obtained from Brute Force simulations for CBG (black bars), original BG (grey bars) and RBG (white bars) schemes. Bars on the left/right correspond to a modest/optimized λ order parameter.

APPENDIX D

ESTIMATION OF THE RATE CONSTANT FOR INITIAL SET OF FFS RUNS

A rate constant of $k_{NL} = (8\ \mu s)^{-1}$ was found from our initial FFS runs, a value in good agreement with the experimental unfolding rate[19] of $k_{NU} = (12.7\ \mu s)^{-1}$, but not with the one obtained by Juraszek and Bolhuis[18] via DFFS ($k_{NL} = (100\ \mu s)^{-1}$). Likely reasons for this discrepancy are discussed below.

Proper sampling of the λ_0 phase space is crucial to obtain a representative TPE. Figure D.1 shows the sampling of selected trajectories from the most common paths followed in $RMSD_{hx}$ vs. n_{wat} (A) and $RMSD_{hx}$ vs. $RMSD_{ca}$ (B) spaces, to be directly compared with plots a,b,d (for Fig. D.1A) and e,f,h (for Fig. D.1B) of Fig. 4 from Ref. 18. Figs. D.1A and D.1B show thorough sampling of the TPE, analogous to that obtained via TPS (plots 4-a and 4-e[18]) and comparable to that found via TIS (plots 4-b and 4-f[18]), but different from the one previously reported using FFS (plots 4-d and 4-h[18]). Fig. D.1B shows three possible routes leading to basin B along a low energy path (discussed in Sec. IVB), with conformations at $\lambda_0 = 0.06$ nm centered around $RMSD_{ca}$ = 0.15, 0.23 and 0.30 nm. On the other hand, Fig 4-h[18] shows conformations at λ_0 that are narrowly distributed along $RMSD_{ca}$ (< 0.17 nm vs. 0.1-0.34 nm for our study), following trajectories similar to those observed in the pathways starting at $RMSD_{ca}$ = 0.15 nm in Fig. D.1A. A similar occurrence is seen in Fig. 4-d (Ref. 18) where only a small region of the n_{wat} phase space at λ_0 is sampled (5-10 vs. 2-15 for our study). Hence, as also indicated by Juraszek and Bolhuis,[18] deficient sampling in their FFS analysis may have resulted mainly from an insufficient collection of λ_0 conformations, limiting access to possible pathways right from the beginning. This is probably attributable to the short length of the MD run (10 ns) they performed to get the λ_0 ensemble and calculating the flux. In addition, only 35 of their trajectories

connected basins A and B (as compared to 394 for our initial FFS runs), limiting the statistics available for calculation of their $P(\lambda_{i+1} \mid \lambda_i)$ values.

The discrepancy between our rate constant and that calculated using FFS in Ref. 18 is also due to the apparent disparity in flux values. Using $\lambda = RMSD_{hx}$, we found $\Phi_{A,0\text{-FFS}} = 2.67\times10^4$ μs^{-1} ($\Phi_{A,0\text{-FFS}}$ represents the flux used for our FFS analysis, calculated from the setup of Sec. IIIB), which is almost four times larger than their flux[18]. This is unexpected given that aside from the simulation time (ours is ten times longer) and the type of thermostat (we used the Andersen thermostat while they used the Nose-Hoover thermostat), the setup of our MD simulations appears to be very similar to the one reported in Ref. 18. To further examine the accuracy of $\Phi_{A,0\text{-FFS}}$, additional fluxes were calculated from three 10 ns simulations using the Nose-Hoover ($\tau = 0.1$ ps), velocity rescaling ($\tau = 0.02$ ps), and conventional Andersen ($\tau = 0.02$ ps) thermostats. The flux values are respectively $\Phi_{A,0\text{-NOSE}} = 9.92\times10^4$ μs^{-1}, $\Phi_{A,0\text{-VRESCALE}} = 2.78\times10^4$ μs^{-1}, and $\Phi_{A,0\text{-ANDERSEN}} = 6.98\times10^4$ μs^{-1}. While a large statistical variation among them is evident, they are in the same order of magnitude as $\Phi_{A,0\text{-FFS}}$. The considerable difference between $\Phi_{A,0\text{-NOSE}}$ and the flux value reported by Juraszek and Bolhuis[18] is unlikely due only to statistical variation but also to differences in other unreported details of their MD simulations.

Finally, another relevant aspect that may lead to divergent results in the rate calculation is the performance of the stochastic thermostat used with FFS. For the case of Ref. 18, a precise comparison between our thermostat and theirs would require further knowledge of their simulation details. Nevertheless, the fact that they estimated a rate constant value of $(1.2\ \mu s)^{-1}$ from their TIS simulations (which used the same thermostat), suggests that enough stochasticity is being imparted to their FFS trajectories. It is also pertinent to note that our experience with FFS shows that proper

selection of the random number generator used for the Andersen thermostat and velocity reassignment is key to successful implementation of this scheme.

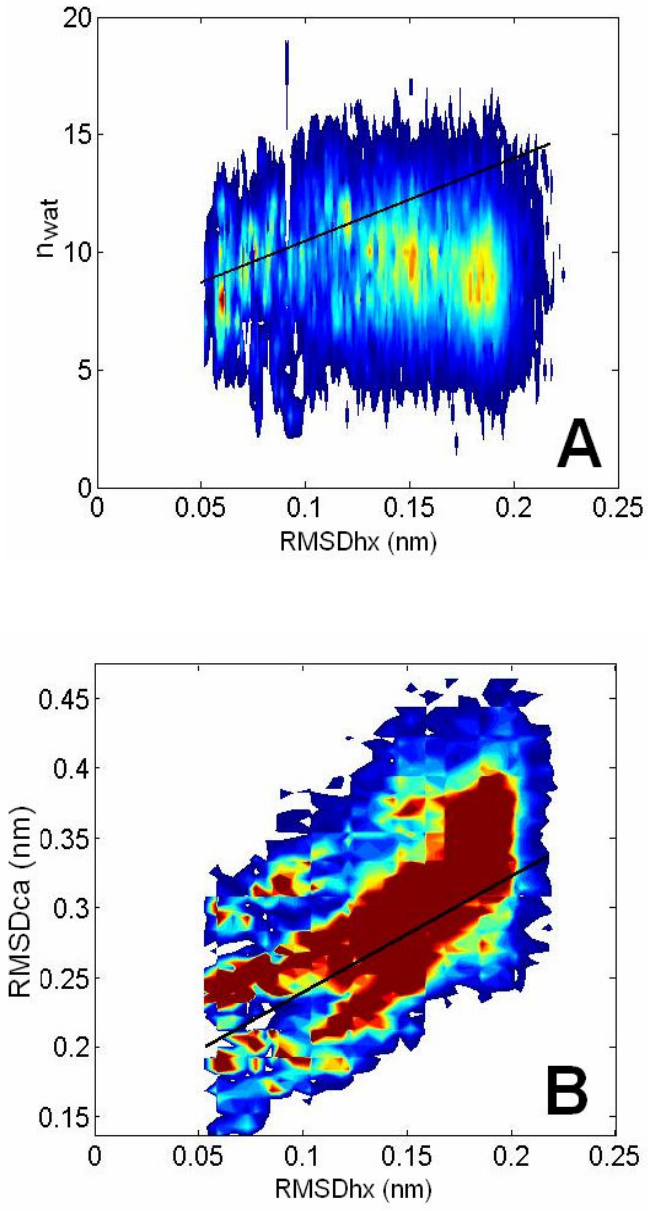


Figure D.1. Sampling of the TPE obtained from our initial set of FFS runs. $RMSD_{hx}$ vs. n_{wat} (*A*) and $RMSD_{ca}$ (*B*). The continuous black lines are included to facilitate comparison with the results obtained by Juraszek and Bolhuis (18).